Package ‘StatR’

April 13, 2024

Version 2.0

Title Estadística Con R

Author Cástor Guisande González, Antonio Vaamonde Liste & Aldo Barreiro Felpeto

Maintainer Cástor Guisande González <castor@uvigo.es>

Description It allows to perform many statistical and graphical analyses.

License GPL (>= 2)

Encoding UTF-8

Depends R (>= 3.1.1)

Repository CRAN

R topics documented:

II1 DESCRIPTIVE STATISTICS

Description

Descriptive statistics, these include measures of position such as arithmetic mean, geometric mean, harmonic mean, weighted average, mode, median, etc., measures of dispersion or sampling variability as variance, standard deviation, coefficient of variation, etc., and measures of distribution as skewness and kurtosis.

Usage

```
II1(data, variables, group=NULL, trim=0.05, quantile=c(0.025,0.05,0.95,0.975),
variablesWM=NULL, variablesW=NULL, groupWM=NULL, file1="Output1.csv",
file2="Output2.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data` Data file.

`variables` Variables for which statistical measures can be calculated.

`group` Variables with grouped data to perform statistical calculations. In case of selecting NULL, no grouping would be carried out and calculations would be made taking into account all the data in each column.

`trim` Cutoff point of the trimmed mean.

`quantile` Quantile values to be calculated.

`variablesWM` If the weighted average interests, the variable or variables for which the weighted average is obtained must be specified. If NULL is selected, the weighted average will not be calculated.

`variablesW` Variable or variables that are used to weight. As before, if NULL is selected, the weighted average will not be calculated.

`groupWM` Variables with grouped data to make weighted average calculations. In case of selecting NULL, any grouping would take place and this would be calculated considering all the data in each columnn.

`file1` CSV FILES. Filename of descriptive statistics.

`file2` CSV FILES. Filename of the weighted average.

`na` CSV FILES. Text that is used in the cells without data.

`dec` CSV FILES. Defining if comma "," or point "." is used as decimal separator.

`row.names` CSV FILES. Logical value that specifies if identifiers are placed in rows or if a vector with a text is placed for each of the rows.

Details

II. DESCRIPTIVE STATISTICS

II.1. MEASURES OF POSITION

The first step in dealing with data is to find some kind of measure that allows us to characterize, differentiate and distinguish the data series. This can be done by determining the position of the data. Within this group are the so-called measures of centrality and other measures that consider different positions of data. Ones or others are used depending on the type of data that you are working with.

II.1.1. Measures of central tendency

II.1.1.1. Arithmetic mean

The arithmetic mean ($\bar{x}$), which is also known simply as mean, or average is calculated using the following formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where *x* are each of the variable values *i* and *n* the number of data.

When the average is a set of all population, it is usually denoted by the symbol μ and, when this is the average of a sample of the population, then used $\bar{x}$.

For grouped data (frequency table) the average is calculated using the following formula:

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n}$$

where *m* and *f* are the mean value and the frequency (number of data) of the class *i*, respectively, and *k* the number of intervals or classes.

The arithmetic mean is the most frequently used because it has a smaller standard error, it is easier to estimate, this tends towards a Normal distribution even if the original data do not present this distribution and, finally, is more sensitive to changes than other measures of position in the distribution of data, which is very important in statistics to determine differences between data series (Sokal & Rohlf, 1981). The problem posed by the arithmetic mean also derives from its sensitivity, which is most affected when rare data is coming out of range.

II.1.1.2. Geometric mean

The geometric mean *GM* is used in some cases with relative data as percentages or growth rates and is calculated using the following formula:

$$GM = \sqrt[n]{x_1 x_2 .... x_n}$$

II.1.1.3. Harmonic mean

The harmonic mean *AM* is used, for example, in some cases where it is necessary to average variations with respect to time. It is calculated as follows:

$$AM = \frac{1}{\frac{1}{n}(\frac{1}{x_1} + \frac{1}{x_2} .... \frac{1}{x_n})}$$

The harmonic mean is always less than or equal to the geometric mean, which in turn is always less than or equal to the arithmetic mean.

II.1.1.4. Mode

Sometimes it is necessary to determine the position of variables that are not numeric but qualitative as, for example, the species that belong to different individuals. In this case we can not use any of the above-mentioned means and, therefore, we must resort to measures like mode. It can also be used for quantitative variables.

The mode of a set of observations is defined as the value that occurs most frequently, i.e., a greater number of times. Mode can be denoted by *Mo*.

If only one value is repeated more times, the distribution will be unimodal, because it only has one mode. In the event that no value is repeated, then by definition, it is considered that there is no mode. It may be the case where two points have the same highest frequency, resulting, in this case, in a bimodal distribution. The multimodal distribution would be in those situations where there are more than two values with the same maximum frequency. Finally, in rare situations in which the distribution of the data is U-shaped, the mid-point of the distribution is called antimode.

II.1.1.5. Median

In a set of observations median is the value arranged in increasing order, half of these are less than this value and half are higher. Be $x_1, x_2, ...x_n$ a ramdon sample of *n* observations sorted increasingly, the median of these data is calculated as follows:

If *n* is an odd number.

$$X_{(\frac{n+1}{2})}$$

If *n* is an even number.

$$\frac{X_{\frac{n}{2}} + X_{(\frac{n}{2}+1)}}{2}$$

The median is often used in data sets which have a very asymmetric distribution since in these cases it is not always suitable to use the arithmetic mean. Finally, it is not affected by rare values that get much out of the normal range, as happens to the arithmetic mean.

II.1.1.6. Trimmed mean

Another measure of central tendency that is often used for exploratory purposes is the trimmed mean, because it is more resistant to extreme values than the average and is not as callous as the median might be. The trimmed mean is defined as the average of the values where a certain percentage of the population on either side of the extreme values has been excluded. For its calculation three steps must be followed: 1) The data is sorted; 2) The percentage of data at each end of the population or sample is discarded and 3) The average is calculated using the remaining data.

This measure serves as an exploratory parameter, it shows the impact of the extreme values on the estimation of averages, because it is calculated at different percentages. It also facilitates the detection of unusual values, which warn researchers of possible data errors that may not be part of the target population of the study or outliers that indicate new directions in research.

II.1.1.7. Weighted average

The weighted average ($\bar{x}_w$) is used in cases where each value is in turn an average. This can also be used when some data is more reliable than other data, simply because some data have been calculated with a greater effort or for any other reason and, therefore, it is necessary to give more weight to these values when estimating the mean. The calculation is based on the following formula:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where *x* are each of the values *i* of the variable, *w* the weight given to the data and *n* the number of data.

An application of the weighted average that deserves special mention, because it can give a lot of information, is to obtain the center of gravity of a distribution of data according to a particular variable. Ecology is a parameter that quantifies the ecological niche of a species, since it provides information on the average value of its distribution for a given variable. For example, one can obtain the temperature, salinity, oxygen concentration, etc. mean that a species may be present. In this case, *w* are each values *i* of the species abundance data, *x* the value of the variable to the value *w* and *n* the number of data..

II.1.2. Other measures of position

Quantile of order *k* is defined as the value of the variable, assuming that this is ranked from least to greatest, that divided it into *k* parts with the same frequency of observations. So there will be *k - 1* quantiles of order *k*.

The first quantile of order *k* lets the *1/k* fraction on the left of the frequency of observations. The second quantile of order *k* lets the *2/k* fraction on the left of the frequency of observations. The rth quantile of order *k* lets the *r/k* fraction on the left of the frequency of observations. For instance, the 15 quantile of order 100 leaves below the 15% of the total value of the full range of values.

The most commonly used quantiles are **percentiles**, **cuartiles** and **deciles**, which are described below.

The **percentiles** are the 99 points that divide the distribution into 100 parts, such that within each 1% of the values of the distribution is included.

The **quartiles** are the three values that divide the variable distribution in 4 equal parts, i.e., in 4 intervals within each 25% of distribution values is included. The 25 percentile *(P_{25})* would be equal to quartile 1 *(Q_1)*, the 50 percentile *(P_{50})* would be equal to quartile 2 (Q_2, also equal to the median), etc.

The **deciles** are the 9 points that divide the distribution into 10 parts, such that within each one is included the 10% of the values of the distribution. The 10 percentil *(P_{10})* would be equal to decil 1 *(D_1)*, The 20 percentil *(P_{20})* would be equal to decil 2 *(D_2)*, etc.

The procedure to find the value of the *j* quantile of *k* order from sorted data from least to greatest, is as follows:

1. Find the *i* position of the *j* th quantile by calculating *nj/k*.

2. If *nj/k* is not an integer, then the *i* position is the next larger integer and the value of the quantile is the data arranged in the position of this larger whole.

3. If *nj/k* is an integer, then the position of the quantile will be *i*=*nj/k* + 0,5 and, thus, the quantile value is the average number of sorted observations *nj/k* y *nj/k* + 1.

II.2. MEASURES OF DISPERSION

In addition to the position, it is also important the dispersion or variability of the data. Two sets of data can have the same mean, but the variability may be different between the two.

Dispersion measures aim to study to what extent, for a given distribution of data, position measurements represent well the distribution set.

For example, if you want to determine whether an arithmetic mean marks a generalized central tendency of behavior of all elements in the set studied, we will have to look at the separation or deviation of each value with respect to the mean. If all values are close to the mean value, it will be representative of them.

That is, the more representative the arithmetic mean of a variable, the more grouped around it are the averaged values and, conversely, it will be more objectionable for not being representative how much greater dispersion of the variable values there is from the average (mean).

Therefore, another type of parameters which measure the dispersion or variability of the data are necessary to complement the information obtained from the mean, as shown below.

II.2.1. Amplitude

The simplest method of estimating the dispersion of the data is by means of the amplitude, also known as range, that is, the difference between the minimum and maximum values of the data series.

II.2.2. Variance and quasi-variance

The best way to measure the dispersion of a set of data is to compare each of them with the mean of the series, and this is exactly what the variance makes (σ^2):

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Often we use the variance of a sample as a proxy or estimate of the unknown variance of the population from which that sample comes. In these cases, the mistake is usually smaller if, instead of considering the variance as an estimator of the sample, is used what is known as quasi-variance (s^2), which is calculated as the previous one, but changing the denominator by *n* - 1:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

In cases where population data is available, *n* should be used and no *n* - 1. However, as these data are a sample of a larger population, the proper way to proceed is to use the quasi-variance instead variance.

Often reference is made to the variance when you are actually calculating the quasi-variance (estimated variance). Most statistics programs only use quasi-variance and not the variance. It is also common to observe that σ^2 and s^2 are used interchangeably to indicate quasi-variance or variance without a defined criterion.

II.2.3. Standard deviation and stardard quasi-deviation

The problem with variance is that to avoid negative values, the differences are squared. In order to have the measure of dispersion in the same units as the mean, the standard deviation is more often used than the variance (σ) which is simply the square root of the variance.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

As is the case with the variance, except in the case that the sample size is identical to the population, it is preferable to use the typical quasi-standard deviation (*s*), instead of the standard deviation. To calculate the quasi-standard deviation, divide by the number of degrees of freedom (*n* - 1) instead of dividing by the total number of data (*n*).

$$s = \sqrt{\frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

It should be mentioned that most of the statistical programs only use quasi-standard deviation and, as the same with variance, it is frequently observed that the standard deviation occurs when the quasi-standard deviation is calculated. It is also often observed that σ and s are used interchangeably for quasi-deviation or standard deviation without a defined criterion.

II.2.4. Absolute deviation with respect to the average and the median

A measure of dispersion that is often used is the mean absolute deviation which is defined as:

$$ADM = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

and the median as:

$$ADMed = \frac{1}{n}\sum_{i=1}^{n}|x_i - Median|$$

This measure indicates the average of the deviations either negative or positive with respect to the overall average and facilitates the interpretation and analysis of the degree of dispersion of the data.

II.2.5. Coefficient of variation

When data sets with different means are compared, the standard deviation does not allow to properly compare which of the two sets of data shows greater variation, since the dataset with higher average also tends to have greater variability. In this case the use of the coefficient of variation is more advisable *CV*, which is calculated as the percentage represented by the standard deviation relative to the average. If the standard deviation is unknown because our data are in a sample, the quasi-standard deviation as an estimator of population standard deviation is used.

$$CV = \frac{s * 100}{\bar{x}}$$

II.2.6. Standard error of the mean

It provides a measure of the accuracy of the estimate of the population mean based on a sample, while the standard deviation measures the variability of the data with respect to the mean in the sample. The standard error is calculated from the standard deviation. When this is known, the quasi-standard deviation is used to obtain the estimated standard error.

$$SE = \frac{s}{\sqrt{n}}$$

II.2.7. Interquartile range

The interquartile range Q is calculated from the 75 (P_{75}) and 25 (P_{25}) percentiles as follows:

$$Q = P_{75} - P_{25} = Q_3 - Q_1$$

This measure of dispersion is used when the central position is expressed by means of the median.

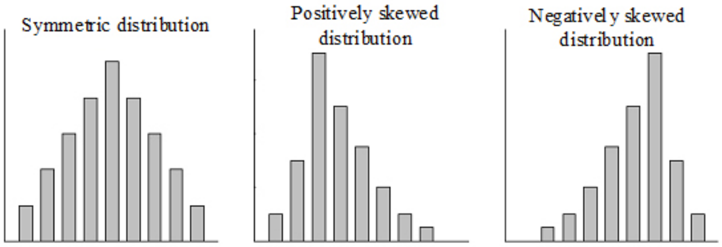
II.3. MEASURES OF DISTRIBUTION

II.3.1. Asymmetry

Measures of asymmetry are intended to determine whether a variable is distributed symmetrically with respect to a central value, or if the distribution data has a different shape on the right side than the left side.

The distribution is symmetric if the right and left side of the central value of the data distribution is the same. The distribution is positively skewed if the higher frequencies are on the left side of the mean, while on the right there are smaller frequencies. Skewness is negative when the frequencies are smaller on the left side (Figure II.1).

Figure II.1. Different types of asymmetry



The arithmetic mean is commonly used as a central reference value, but also the median can be used.

There are many ways of measuring the symmetry and one of them is the formula that is shown below (Fisher asymmetry coefficient) which uses the arithmetic average as a central value:

$$A = \frac{n\left(\sum_{i=1}^{n}(x_i - \bar{x})^3\right)}{(n-1)(n-2)\sigma^3}$$

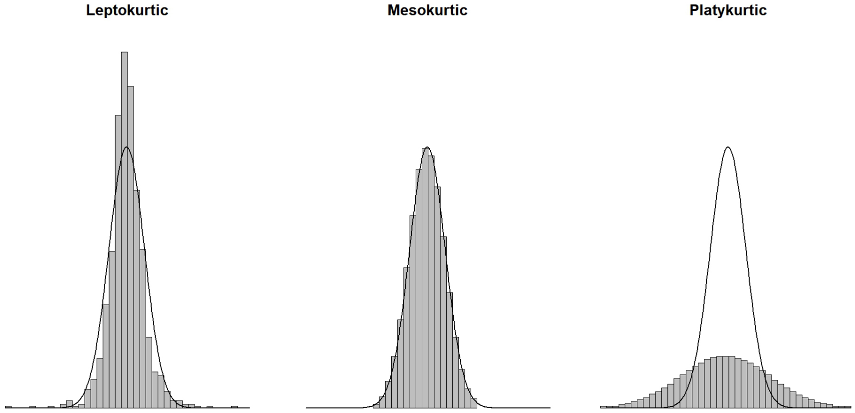
Being σ the estimated standard deviation or sample quasi-standard deviation. When the value is close to zero the distribution is symmetric, the asymmetry is positive when the value is greater than zero and negative when the value is less than zero.

II.3.2. Kurtosis

Kurtosis measure characterizes the tails (or outlier potential) of distribution of the data. Historically, it was thought that higher kurtosis implies a more or less "pointed" distribution. For this reason, kurtosis was historically called a measure of "peakedness", "pointiness", or "central concentration". However, the notion that kurtosis somehow measures "peakedness" (flatness, pointiness or modality) is incorrect. Kurtosis is not informative about the shape of the peak - its only unambiguous interpretation is in terms of tail extremity; i.e., either existing outliers (for the sample kurtosis) or propensity to produce outliers (for the kurtosis of a probability distribution) (see Westfall, 2014).

When the kurtosis value is positive, the distribution is leptokurtic, meaning that the outlier character of the distribution is more extreme than that of a normal distribution (Figure II.2, left panel). When the kurtosis value is zero or close to zero, the distribution has similar outlier characteristic as the Normal distribution, and is called mesokurtic (Figure II.2, central panel). Finally, when the kurtosis value is negative, the distribution is platykurtic, meaning that the outlier character of the distribution is less extreme than that of a normal distribution (Figure II.2, right panel)

Figure II.2. Types of kurtosis (bars) compared to the Normal distribution (solid line).



As with asymmetry, there are several ways to estimate kurtosis of a data distribution, but one of the most used is Fisher's correlation coefficient as shown below:

$$C = \frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)(n-1)}{(n-1)(n-2)(n-3)\sigma^4}$$

EXAMPLE

Position measurements such as average arithmetic, geometric mean, harmonic mean, median, etc., and measures of dispersion or sampling variability such as variance, standard deviation, coefficient of variation, etc., are estimated from the abundance of several fish species in two zones (S1 and S2). These parameters were estimated by performing the calculations for each family and genus. The weighted average of the temperature for the two sampling sites based on the abundance of the species is also calculated.

Value

A CSV file is exported with position measurements such as arithmetic mean, geometric mean, harmonic mean, mode, median, etc. and measures of dispersion or sampling variability as variance, standard deviation, coefficient variation, etc. Optionally, a second CSV file, that includes the weighted average, is exported.

References

Sokal RR & Rohlf FJ (1981) *Biometry*. WH Freeman and Company, New York.

Westfall, R.H. (2014) Kurtosis as Peakedness, 1905. *RIP* (2014) *American Statistician*, 68: 191-195.

Examples

```
data(ZII1)

II1(data=ZII1, variables=c("S1","S2"),group=c("Family","Genus"),
variablesWM=c("Temperature"), variablesW=c("S1","S2") )
```

III1 *NORMAL DISTRIBUTION*

Description

It calculates the probability of a Normal distribution indicating the mean and standard deviation.

Usage

```
III1(value, mean, sd, prob.g=FALSE)
```

Arguments

`value`	Value for which the probability of the Normal distribution will be calculated.
`mean`	Mean of the Normal distribution.
`sd`	Standard deviation of the Normal distribution.
`prob.g`	It indicates if you want to get the greater TRUE or lesser probability FALSE of the value.

Details

III. DISTRIBUTION

Another important information is related to the distribution that a variable has. It may be the case that two variables have exactly the same mean and dispersion, but different types of distribution. Therefore, in addition to the information about the position and dispersion measures, which is explained in function II1, it is necessary to know the distribution of the values of the variable. For studying the distribution of a variable, and even compare averages and variances between variables, what is done is to compare the frequencies of the values of the variable with the probabilities resulting from theoretical models of distributions. The theoretical distribution model used varies depending on the variable on which we are working.

III.1. TYPES OF VARIABLES

There are basically two types of variables: *qualitative* and *quantitative*. The first are not expressed numerically (gender, species to which an individual belongs, province of birth) while numeric codes can be used to represent their values (for example in the variable gender, instead of male and female can be appointed as 1 and 2). The quantitative is directly expressed in numerical terms (number of leaves of a plant, age, length, temperature, etc.)

Qualitative variables can in turn be of two types: *nominal*, if values are not ordered in a natural way (birthplace, species), and ordinal if values have a rank (order) (for example a variable toxicity that takes values such as nothing, little, pretty and very toxic).

Quantitative variables can also be of two types: *discrete* and *continuous*. They are discrete variables when they can only take specific values, and when any value is possible between two consecutive ones (number of children that can have a family, number of leaves of a plant, etc.). In the case of discrete variables, the corresponding probability is assigned to every value of the variable on which the number of times that value is repeated will depend in relation to the remaining values. A correspondence between values and their respective probabilities is called *probability function*.

Continuous quantitative variables are those that can take any value along a continuum, so there are no consecutive values, as between any two values remain infinite possible values (temperature, length). Continuous variables can be grouped into categories, but in an arbitrary way. For example, the variable height can be divided into categories as small, normal and high, and the limits of each of these categories can be set arbitrarily. Unlike discrete variables, for continuous variables it is not useful to establish the correspondence between values and probabilities. What is done is to calculate the probability contained in a given segment or range of values that divided by the amplitude of the segment is the mean probability density of the segment, from which the probability density for each value is determined. The correspondence between the values and their respective probability density is called *density function*.

The distinction between these four types of variables is important for several reasons:

1. The calculation of some dispersion or position measures does not make sense with qualitative variables, for example in the case of the gender variable.

2. For the correct application of techniques of statistical analysis: the majority of non-parametric tests require that the variable be at least ordinal, and many methods of multivariate analysis require variables be quantitative (e.g. factor analysis or discriminant analysis).

III.2. DISTRIBUTIONS FOR CONTINUOUS VARIABLES

III.2.1. Normal

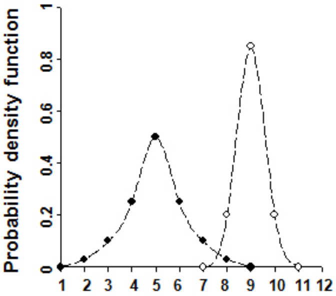
The normal distribution is one of the most important because it is seen in many series of data and, also, different types of statistical analyzes have as a condition for being applied to the data series, that this type of distribution exists. Several mathematicians took part in its development, including the mathematician and astronomer Carl Friedrich Gauss (1777-1885), so it is sometimes called "bell shaped curve" or Gaussian distribution in his honor. The density function of the Normal distribution is described by using the following formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}$$

where *f(x)* is the probability density of the value *x*, μ is the mean and σ is the standard deviation.

The shape of the normal distribution varies depending on the media μ and the standard deviation σ as shown in Figure III.1.

Figure III.1. Normal distributions with different means and standard deviations.



III.2.1.1. Applications of the Normal distribution

The table of the Normal distribution N(0,1), that is to say with $\mu = 0$ and $\sigma = 1$ (Table 1 of Appendix I in Guisande et al., 2011), allows to calculate probabilities relating to any other normal distribution with different μ and σ. To do this, simply define the variable, that is, calculate the value *Z* (deviation units compared with the average) corresponding to the values *x* indicated by the operation:

$$Z = \left(\frac{x-\mu}{\sigma}\right)$$

This value Z which is calculated from the variable X allows us to obtain the probabilities corresponding to any interval (see example) in the tables.

III.2.2. t of Student

When a variable follows a Normal distribution, the mean of a random sample of that variable also has a Normal distribution, and its mean is the unknown population mean μ. This can be used to estimate μ. However, the standard deviation of the population is not often known σ (it only works with a sample of individuals of the total population) and, in addition, it may happen that the number of observations in the sample is small (less than 30). In these cases, you can use the quasi-standard deviation of the sample (*S*) together with the distribution *t* of Student:

$$t = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}}$$

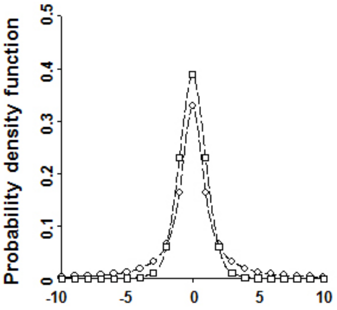
The probability density function of the distribution *t* of Student is given by the following expression:

$$f(x) = \frac{1}{\sqrt{\upsilon\pi}} \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\Gamma\frac{\upsilon}{2}} \left(1+\frac{x^2}{\upsilon}\right)^{-\left(\frac{\upsilon+1}{2}\right)}$$

The *t* Student distribution can have different forms depending on the degrees of freedom (υ) (Figure III.2). The general appearance of the distribution *t* is similar to the standard Normal distribution. However, the distribution *t* has broader queues than the Normal one, that is to say, the probability of the queues is greater than the Normal distribution. The distribution *t* is transformed into a Normal distribution when the number of data tends to infinity. The critical values for different levels of significance and different degrees of freedom can be seen in Table 2 of Appendix I in Guisande et al. (2011).

Figure III.2. The distribution density functions of Student

for 1 (circle) and 10 (square) degrees of freedom.



The main applications of the distribution *t* of Student in statistical inference are: 1) To estimate, using confidence intervals, the population mean and 2) Estimating and testing hypotheses on a mean difference.

Hypotheses or assumptions to apply the *t* of Student are in each group that the studied variable follows a Normal distribution and that dispersion in both groups should be homogeneous (hypothesis of homoscedasticity = equality of variances) although this statistic can also be used without assuming equality of variances.

III.2.3. Chi-square

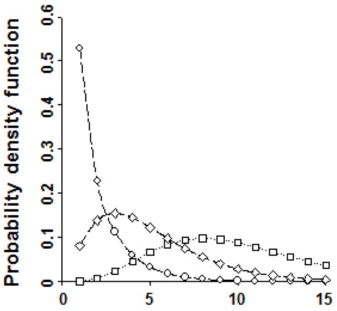
The density function of the Chi-square (χ^2)distribution is described by the following expression:

$$f(x) = \frac{1}{2^{\frac{\upsilon}{2}}\Gamma\left(\frac{\upsilon}{2}\right)} x^{\left(\frac{\upsilon}{2}-1\right)} e^{-\frac{x}{2}}$$

where υ are the degrees of freedom and *x* is not negative.

Unlike the Normal distribution, since the distribution χ^2 depends on the degrees of freedom, there is no typical curve but the distribution χ^2 can have different shapes depending on the degrees of freedom (Figure III.3).

Figure III.3. The distribution density functions of χ^2 for 1 (circle), 5 (rhombus) y 10 (square) degrees of freedom.



The value of the variable that leaves on its right an area α under the density curve is called the corresponding critical point to the level of significance α. The critical points of different levels of significance and different degrees of freedom are tabulated (Table 3 of Appendix I Guisande et al., 2011).

There are three main applications that have the distribution χ^2: The goodness-of-fit test, the test of independence and the homogeneity test. Moreover, this distribution plays an important role in many other statistical tests.

1. The goodness-of-fit test is the approach of how a sample can be considered as belonging to a population with a known theoretical distribution. It is a method frequently used to determine if a data set has a Normal distribution, Poisson, etc.

2. The test of independence determines if two characters *X* and *Y* of a population are dependent or independent. For example, to determine if the survival of the descendants of the females in a population is independent or dependent of the daily amount of food that females receive.

3. The homogeneity test allows to determine if multiple samples that are studying the same character *A* have been taken or not in the same population, with respect to the *A* feature. For example, several groups of individuals in a population who have undergone the same dose of different metals have been selected and to determine if metals affect the survival of individuals differently.

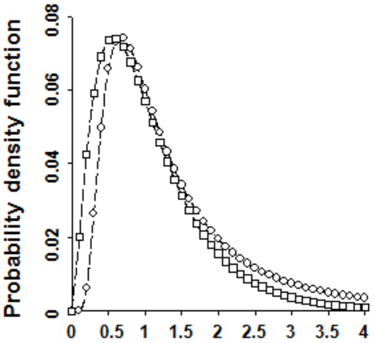
III.2.4. *F* of Fisher-Snedecor

The probability density function of the distribution *F* Fisher-Snedecor is given by the following expression:

$$f(x) = \frac{\Gamma\left(\frac{\upsilon+\omega}{2}\right)}{\Gamma\frac{\upsilon}{2}\Gamma\frac{\omega}{2}}\upsilon^{\frac{\upsilon}{2}}\omega^{\frac{\omega}{2}}\frac{x^{\frac{\upsilon}{2}-1}}{(\omega+x\upsilon)^{\frac{\upsilon+\omega}{2}}}$$

where υ y ω are the degrees of freedom of the numerator and denominator respectively, being x nonnegative. By relying on two types of degrees of freedom, the density function can have different forms (Figure III.4).

Figure III.4. The distribution density functions of*F* Fisher-Snedecor for different degrees of freedom$F_{(30,5)}$ (circle) and $F_{(5,30)}$ (square).



The critical values of the distribution *F* Fisher-Snedecor with different levels of significance and degrees of freedom are shown in Table 4 of Appendix I Guisande et al. (2011).

This distribution is mainly used in two types of situations, requiring in both cases that the distribution of the variables be Normal:

1. To test whether two samples come from populations with equal variances. This is a useful test to determine if a Normal population has a greater variation than the other and it is important because, when it compares averages, several statisticians present the homogeneity of variances as requirement.

2. It also applies when it comes to simultaneously compare several population means.

EXAMPLES

The body length of a species in a given population is distributed according to Normal μ = 10.8 cm and σ = 3.7 cm.

Case 1. Calculate the probability that an individual may have a size of 8.9 cm.

Result: 0.3038

Case 2. Calculate the probability that an individual may have a size between 8.9 and 12.4 cm.

Result: 0.3635

Value

The probability value for the specified mean and deviation, is obtained considering a Normal distribution.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Diaz de Santos, Madrid, 978 pp.

Examples

```
#Case 1
III1(value=8.9, mean=10.8, sd=3.7, prob.g=FALSE)

#Case 2

III1(value=12.4, mean=10.8, sd=3.7, prob.g=FALSE)-III1(value=8.9,
mean=10.8, sd=3.7, prob.g=FALSE)
```

III2 *DISTRIBUCION BINOMIAL*

Description

Calculate the probability of a Binomial Distribution.

Usage

```
III2(n=NULL, x=NULL, p=NULL, option=1)
```

Arguments

n	Total number of individuals.
x	Number of individuals.
p	Probability of success.
option	1. Exactly the probability p(X=x). 2. Probability that is greater or equal than p(X>=x). 3. Probability that is strictly greater p(X>x). 4. Probability that is less or equal than p(X<=x). 5. Probability that is strictly less p(X<x).

Details

III. DISTRIBUTION

III.3. DISTRIBUTIONS FOR DISCRETE VARIABLES

III.3.1. Binomial distribution

A variable has a binomial distribution when only two possible outcomes are possible: «success» and «failure» , being the probability of each of them constant in a series of repetitions, that is to say, nor the probability of success or failure of change from one test to another, and in addition each test result is independent of the outcome of the other tests. The probability of success is represented by *p* and the probability of failure is represented by *q* = 1 - *p*.

Discrete variables instead of the density function use the function of probability or amount, giving a probability for each value of the variable. The binomial probability function is expressed by the following formula:

$$f(x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}$$

EXAMPLES

The application of a specific treatment to a group of individuals presenting a parasite provides an improvement in a 67 % of the cases. If treatment is applied to 8 individuals:

The value of *p* = 0.67 and, therefore, the value of *q* is 0.33.

Case 1. What is the probability that 7 individuals improve in a sample of 8?

$$\frac{8!}{7!(8-7)!}0.67^7 0.33^{8-7} = 0.16$$

Result: 0.16

Case 2. What is the probability that at least 3 individuals improved?

The probability that at least three individuals improve will be 1 minus the probability that two individuals improve minus the probability that one individual improves minus the probability that anyone improves:

$$1 - \frac{8!}{2!(8-2)!}0.67^2 0.33^{8-2} - \frac{8!}{1!(8-1)!}0.67^1 0.33^{8-1} - \frac{8!}{0!(8-0)!}0.67^0 0.33^{8-0} = 0.981$$

Result: 0.981

Value

The probability value is obtained considering a Binomial Distribution.

Examples

```
#Case 1
III2(n=8, x=7, p=0.67, option=1)

#Case 2
1-III2(n=8, x=2, p=0.67, option=1)-III2(n=8, x=1, p=0.67, option=1)- III2(n=8,
x=0, p=0.67, option=1)
```

III3 *HYPERGEOMETRIC DISTRIBUTION*

Description

Calculate the probability of a hypergeometric distribution.

Usage

```
III3(N=NULL, N.p=NULL, n=NULL, x=NULL, option=1)
```

Arguments

`N`	Total number of elements in the population.
`N.p`	Value of Np, number of elements with initial probability p.
`n`	Sample size.
`x`	Value of the random variable for which the probability is calculated.
`option`	1. Exactly the probability p(X=x). 2. Probability that is greater or equal than p(X>=x). 3. Probability that is strictly greater than p(X>x). 4. Probability that is less or equal than p(X<=x). 5. Probability that is strictly less than p(X<x).

Details

III. DISTRIBUTION

III.3. DISTRIBUTIONS FOR DISCRETE VARIABLES

III.3.2. Hypergeometric distribution

In the hypergeometric distribution the variable is also random and dichotomous as in the binomial distribution, but differs from the latter in two important characteristics: the population is finite, while the binomial can be infinite and, in addition, the probability *p* changes, it is not constant, because the result of each test depends on the result of the previous ones. The probability function is expressed by the following formula:

$$f(x) = \frac{\frac{N_p!}{x!(N_p-x)!}\frac{N_q!}{(n-x)!(N_q-n+x)!}}{\frac{N!}{n!(N-n)!}}$$

where N_p and N_q are the number of elements with initial probability p and q, respectively (i.e., p = N_p/N and q = N_q/N), N the total number of elements and n the number of elements in the sample extracted from the N of the population.

EXAMPLES

It has been proven that in a batch of 30 vaccines, 8 are in poor condition. 4 vaccines of the lot have already been used.

Case 1. What is the probability that at least one of the vaccines supplied is in poor condition?

The probability that none of the vaccines supplied is in poor condition is calculated

$$1 - \frac{\frac{22!}{4!(22-4)!}\frac{8!}{(4-4)!(8-4+4)!}}{\frac{30!}{4!(30-4)!}} = 0.733$$

Result: 0.733

Case 2. What is the probability that 3 of the vaccines supplied are in poor condition?

$$\frac{\frac{22!}{1!(22-1)!}\frac{8!}{(4-1)!(8-4+1)!}}{\frac{30!}{4!(30-4)!}} = 0.045$$

Result: 0.045

Value

The probability value is obtained by considering a hypergeometric distribution.

Examples

```
#Case 1
1- III3(N=30, N.p=8, n=4, x=0, option=1)

#Case 2
III3(N=30, N.p=8, n=4, x=3, option=1)
```

III4 *POISSON DISTRIBUTION*

Description

Calculate the probability of a Poisson distribution.

Usage

```
III4(lambda=NULL, x=NULL, option=1)
```

Arguments

`lambda`	Value of the lambda parameter.
`x`	Value on which the probability will be calculated.
`option`	1. Exactly the probability of p(X=x). 2. Probability that is greater or equal than p(X>=x). 3. Probability that is strictly greater than p(X>x). 4. Probability that is less or equal than p(X<=x). 5. Probability that is strictly less than p(X<x).

Details

III. DISTRIBUTION

III.3. DISTRIBUTIONS FOR DISCRETE VARIABLES

III.3.3. Poisson Distribution

A Poisson process is a process of independent events that is characterized by:

1. The number of events in two different intervals is always independent.

2. The probability of an event occurring in an infinitesimal interval is proportional to the length of the interval.

3. The probability that more than one event occur in a very small interval *h* is 0.

4. The events are expressed per unit of area, time, etc

The Poisson distribution describes the number of events in a time unit of a Poisson process. Many phenomena are modeled as a Poisson process, for example the number of accidents in a particular area of a road.

The most important differences with respect to the binomial distribution are that this distribution is applied to events that may have a very low probability and, in addition, the size of *n* is infinite. In some cases the Poisson distribution is used as an approximation to the binomial when *n* is very large and therefore, it is difficult to calculate the binomial and also when the probability of any of the events is very low. The probability function of the Poisson distribution is expressed by the following formula:

$$f(x) = \frac{\lambda^x}{x!e^\lambda}$$

where λ is the mean or average of events per unit of time and *x* is the variable that indicates the number of events.

EXAMPLES

Case 1. The mean species abundance is 23 individuals per 100 m^2. As these are events per unit area a Poisson is used. What is the probability of not finding any individual in 25 m^2?

$$\lambda = \frac{23individuals * 25m^2}{100m^2} = 5.75$$

$$\frac{5,75^0}{0!e^{5,75}} = 0.0031$$

Result: 0.0031

Case 2. The mean number of observed tigers is 120 in 30 days. As these are events per unit time a Poisson is used. What is the probability of seeing 5 tigers in 10 days?

$$\lambda = \frac{120tigres * 10d}{30d} = 40$$

$$\frac{40^5}{5!e^{40}} = 3.63^{-12}$$

Result: 3.63^{-12}

Value

The probability value is obtained considering a Poisson distribution.

Examples

```
#Case 1
III4(lambda=5.75, x=0, option=1)

#Case 2
III4(lambda=40, x=5, option=1)
```

IV1 *CONFIDENCE INTERVAL FOR THE MEAN OF A NORMAL POPULATION*

Description

Calculate the confidence interval for the mean of a Normal population.

Usage

```
IV1(data, variables, group=NULL, alfa=0.05, total=FALSE, file="Output.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Data file.
`variables`	Variable or variables for which the interval confidence is calculated.
`group`	Variables that gather data for the calculations. In case of selecting NULL, there would be no grouping and it would be calculated considering all the data of the selected variables.
`alfa`	Confidence level.
`total`	Logical value that if FALSE means that the entire population has not been sampled, which will be in most of the studies, and if TRUE means that the entire population has been sampled.
`file`	CSV FILE. Output file name.
`na`	CSV FILE. Text that is used in cells without data.
`dec`	CSV FILE. Defining if comma "," or point "." is used as decimal separator.
`row.names`	CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

IV. CONFIDENCE INTERVALS

A way to estimate certain parameters, such as different types of mean, variance, standard deviation, etc., that allow us to have information about the variables we are studying was described in the function II1. In most cases, these parameters are estimated from the data of a sample, and not from the whole population. To measure the reliability of the estimate, it is necessary to estimate the confidence interval of the parameter.

The confidence interval of a parameter to the confidence level 1 - α is one that satisfies the property from the probability of their ends taking such values and which the parameter between them is equal to 1 - α (Viedma, 1989). The ends of the confidence interval of a parameter are called confidence limits.

IV.1. CONFIDENCE INTERVAL FOR THE MEAN OF A NORMAL POPULATION

IV.1.1. Known standard deviation

On the assumption that the standard deviation of the population is known, the confidence interval of the mean at the level of $1 - \alpha \left(I_{\mu}^{1-\alpha}\right)$ is calculated by the following interval estimation:

$$I_{\mu}^{1-\alpha} = \left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

The critical values used are the areas under the standard Normal curve (Table 1 Appendix I Guisande et al., 2011).

IV.1.2. Unknown standard deviation

The most common is to work with a few elements of the total population. This means that we do not know the standard deviation of the entire population. In this type of situation, the confidence interval is calculated in two different ways depending on the size of the sample.

IV.1.2.1. Small sample size (<30)

If the sample size is small, instead of using the Normal distribution and, therefore, the statistical *Z*, what is used is the distribution *t* of Student with *n* - 1 degrees of freedom. The confidence interval of the mean level of confidence $1-\alpha \left(I_{\mu}^{1-\alpha}\right)$ is:

$$I_{\mu}^{1-\alpha} = \left(\bar{x} - t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}\right)$$

The critical values of the *t* Student distribution are used (Table 2 Appendix I Guisande et al., 2011).

IV.1.2.2. Large sample size (≥30)

When the sample size tends to infinity, the *t* of student tends to a standard Normal distribution. For this reason, most of the tables of the distribution *t* have only values corresponding to a number of degrees of freedom between 1 and 30; when this is higher, the table of the Normal is directly used as approximation, since differences obtained are very small.

The statistical programs do not need to use this Normal approach, because they can calculate the exact value of *t* whatever the number of degrees of freedom. The differences observed between the intervals obtained with STATISTICA or SPSS and this function are due to the statistical programs which do not take into account if the number of data is small (< 30) or large ($\geq$ 30), since they are always working with the distribution *t* of Student and never use the Normal approach.

If the sample size is large, the same expression as described above for the case of known standard deviation is used, but with the difference that the quasi-standard deviation sampling (*s*) is used instead of the population standard deviation (σ).

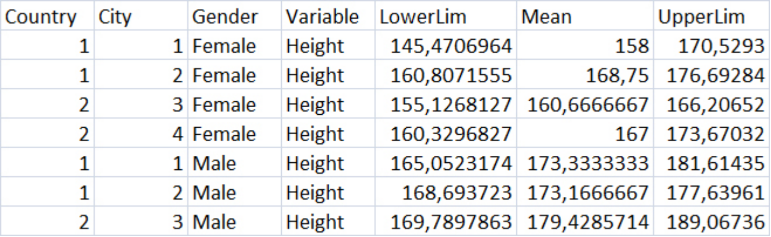
$$I_{\mu}^{1-\alpha} = \left(\bar{x} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$$

The critical values used are the areas under the standard Normal curve (Table 1 Appendix I Guisande et al., 2011).

EXAMPLE

Data on height and weight in men and women from different countries and cities. It is intended to calculate confidence intervals to 95% of the height and weight for both men and women. Further calculations are done for each country and city.

The results obtained are shown in the following table, indicating the estimated value in each group, and the confidence limits to 95% for the corresponding population mean. It is assumed that the variable has Normal distribution, and that the population standard deviation is unknown and must be estimated using the sampling data.



Country	City	Gender	Variable	LowerLim	Mean	UpperLim
1	1	Female	Height	145,4706964	158	170,5293
1	2	Female	Height	160,8071555	168,75	176,69284
2	3	Female	Height	155,1268127	160,6666667	166,20652
2	4	Female	Height	160,3296827	167	173,67032
1	1	Male	Height	165,0523174	173,3333333	181,61435
1	2	Male	Height	168,693723	173,1666667	177,63961
2	3	Male	Height	169,7897863	179,4285714	189,06736

Value

A CSV file is obtained with the lower and upper range limits of the confidence interval and the mean value.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Viedma, J.A. (1989) Bioestadística. *Métodos en Medicina y Biología.* Librería Universitaria, Alicante.

Examples

```
## Not run:

data(ZIV1)

IV1(data=ZIV1, variables=c("Height","Weight"),
group=c("Country","City","Gender"), alfa=0.05, total=FALSE)


## End(Not run)
```

IV2 *CONFIDENCE INTERVAL OF THE DIFFERENCE BETWEEN NORMAL POPULATION MEANS*

Description

Calculate the confidence interval of the difference between Normal population means.

Usage

```
IV2(data, variables, varSel, group=NULL, alfa=0.05,  file="Output.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Data file.
`variables`	Variable or variables for which the confidence interval will be calculated.
`varSel`	It specifies the variable or variables that two studied groups have. It is important to mention that if the variable was previously defined as grouping variable, it can not be used in this section. If the variable has more than two groups, it is necessary to specify which are the groups that are going to be studied. In the example, the confidence interval on the difference between means of the City 1 and City 2 was calculated. If the variable has only 2 groups, as for example «sex», it is not necessary to specify the groups to analyze.
`group`	Variables that gather data for calculations. In case of selecting NULL, there would be no grouping and it would be calculated considering all the data of the selected variables.
`alfa`	Confidence level.

`file` CSV FILE. Output file name.

`na` CSV FILE. Text used in cells without data.

`dec` CSV FILE. Defines whether the comma "," or dot "." is used as decimal separator.

`row.names` CSV FILE. Logical value that specifies whether identifiers are placed in rows or a vector with a text for each of the rows.

Details

IV. CONFIDENCE INTERVALS

IV.2. CONFIDENCE INTERVAL FOR THE DIFFERENCE IN TWO NORMAL POPULATION MEANS

When comparing means of different populations, it is important to estimate the confidence interval for the difference between means. As with the range that was analyzed in function IV1, the calculation is different depending on whether the population standard deviation (known variance) is known or should be estimated with sampling data (unknown variance).

IV.2.1. Known variances

The confidence interval $\mu_1 - \mu_2$ to confidence level $1 - \alpha \left(I^{1-\alpha}_{\mu_1-\mu_2}\right)$ is calculated by using the following expression:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

The critical values used are the areas under the standard Normal curve (Table 1 Appendix I Guisande et al., 2011).

IV.2.2. Unknown variances

In the event that two means are compared, being the independent variables with Normal distribution, but the population variances are unknown, other different expressions depending on the size of the sample must be applied.

IV.2.2.1. Large sample size (≥30)

For large samples, the confidence interval $\mu_1 - \mu_2$ to confidence level $1 - \alpha \left(I^{1-\alpha}_{\mu_1-\mu_2}\right)$ is calculated by using the following expression:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

As seen, the only difference from the situation in which the variances are known, is that the quasi-variance is used instead of variance.

The critical values used are the areas under the standard Normal curve (Table 1 Appendix I Guisande et al., 2011).

IV.2.2.2. Small sample size (<30)

There are two possible situations: equal or unequal variances.

IV.2.2.2.1. Equal variances

Assuming that the variances are unknown, but knowing that there are no significant differences between them, in this case the distribution *t* of Student is used and the confidence interval $\mu_1 - \mu_2$ to the confidence level $1 - \alpha \left(I^{1-\alpha}_{\mu_1 - \mu_2} \right)$ is calculated by means of the following expression:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[\begin{array}{l} (\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2};n_1+n_2-2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}, \\ (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2};n_1+n_2-2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \end{array} \right]$$

The critical values of the distribution *t* of Student are used (Table 2 Appendix I Guisande et al., 2011).

IV.2.2.2.2. Unequal variances

In those cases where the variances are unknown and, furthermore, there are significant differences between them, or are simply not known whether the variances are equal or not, the interval $\mu_1 - \mu_2$ al nivel de confianza $1 - \alpha \left(I^{1-\alpha}_{\mu_1 - \mu_2} \right)$ is calculated as follows:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2},f}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2},f}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

where *f* are the degrees of freedom, called approximation of Welch, which is calculated as follows, taking *f* the value of the nearest integer:

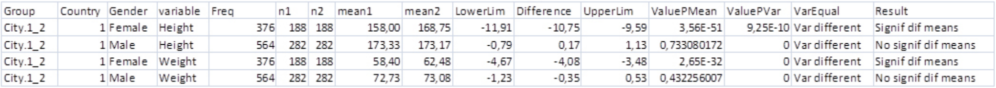
$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1+1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2+1}} - 2$$

The critical values of the distribution *t* of Student are used (Table 2 Appendix I Guisande et al., 2011).

EXAMPLE

Data of height and weight in men and women from different countries and cities. The objective is to estimate the confidence interval for the difference between means in weight and height, in men and women, between cities. Calculations are performed for each country and by gender.

The results obtained are shown in the following table, which shows, for the selection made, the sample sizes, the estimated mean difference and the confidence interval at 95% for the difference between population means, and the P values of the contrast of invalidity of the mean difference (a value less than 0.05 evidences that the means are different), and contrast of equal variances (P value greater than 0.05 can accept that the variances are equal in both groups). In all cases it is assumed that the variable has Normal distribution and that the population variance is unknown, so it is estimated with the sample data.



Group	Country	Gender	variable	Freq	n1	n2	mean1	mean2	LowerLim	Difference	UpperLim	ValuePMean	ValuePVar	VarEqual	Result
City.1_2	1	Female	Height	376	188	188	158,00	168,75	-11,91	-10,75	-9,59	3,56E-51	9,25E-10	Var different	Signif dif means
City.1_2	1	Male	Height	564	282	282	173,33	173,17	-0,79	0,17	1,13	0,733080172	0	Var different	No signif dif means
City.1_2	1	Female	Weight	376	188	188	58,40	62,48	-4,67	-4,08	-3,48	2,65E-32	0	Var different	Signif dif means
City.1_2	1	Male	Weight	564	282	282	72,73	73,08	-1,23	-0,35	0,53	0,432256007	0	Var different	No signif dif means

Value

A CSV file is obtained with the lower and upper limits of the confidence interval of the difference between means of Normal populations.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Examples

```
## Not run:

data(ZIV1)

IV2(data=ZIV1, variables=c("Height","Weight"),
group=c("Country","Gender"), alfa=0.05, varSel=list(c("City","1","2")))


## End(Not run)
```

IV3 *CONFIDENCE INTERVAL OF THE VARIANCE AND STANDARD DEVIATION OF A NORMAL POPULATION*

Description

It calculates the confidence interval of the variance and standard deviation of a Normal population.

Usage

```
IV3(data, variables, group=NULL, alfa=0.05, IntervType="sd", file="Output.csv",
na="NA", dec= ",", row.names=FALSE)
```

Arguments

`data` Data file.

`variables` Variable or variables whose variance will be used to calculate the confidence interval.

`group` Variables that gather data for calculations. In case of selecting NULL, there would be no grouping and it would be calculated considering all the data of the selected variables.

`alfa` Confidence level.

`IntervType` It specifies if the confidence interval of the variance "Variance" or standard deviation "sd" is performed.

`file` CSV FILE. Output file name.

`na`	CSV FILE. Text used in cells without data.
`dec`	CSV FILE. Defines whether the comma "," or dot "." is used as decimal separator.
`row.names`	CSV FILE. Logical value that specifies whether identifiers are placed in rows or a vector with a text for each of the rows.

Details

IV. CONFIDENCE INTERVALS

IV.3. CONFIDENCE INTERVAL OF THE VARIANCE AND STANDARD DEVIATION OF A NORMAL POPULATION

Sometimes it is important to study the greater or lesser concentration of values around the mean. For example, the time that an individual remains immune after being vaccinated, it is not only important to know the average length, but also that the variability of the duration of the effect will not be very large from one individual to another. For this reason the variance or standard deviation can be estimated using a confidence interval.

The interval of the variance at a confidence level $1-\alpha \left(I_{\sigma^2}^{1-\alpha}\right)$ is calculated by the following interval estimator, where, s^2 is the quasi-variance of the sample:

$$I_{\sigma^2}^{1-\alpha} = \left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \right)$$

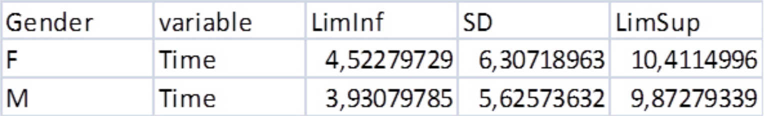
The confidence interval of the standard deviation at a confidence level $1-\alpha \left(I_{\sigma}^{1-\alpha}\right)$ is estimated using the following interval estimator:

$$I_{\sigma}^{1-\alpha} = \left(\frac{\sqrt{n-1}s}{\sqrt{\chi^2_{\frac{\alpha}{2};n-1}}}, \frac{\sqrt{n-1}s}{\sqrt{\chi^2_{1-\frac{\alpha}{2};n-1}}} \right)$$

The critical values of the distribution χ^2 are used. (Table 3 Appendix I Guisande et al., 2011).

EXAMPLE

Time data (in months) that a group of women and men remain immune after they are given a vaccine. It is necessary to know a confidence interval at 95% of the standard deviation of women and men. The results obtained are shown in the table below, which shows the estimated standard deviation along with its 95%, confidence limits, assuming that the variable has a Normal distribution.



Gender	variable	LimInf	SD	LimSup
F	Time	4,52279729	6,30718963	10,4114996
M	Time	3,93079785	5,62573632	9,87279339

Value

A CSV file is obtained with the lower and upper limits of the confidence interval for the variance or standard deviation of a Normal population.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Examples

```
## Not run:

data(ZIV3)

IV3(data=ZIV3, variables=c("Time"), group=c("Gender"))


## End(Not run)
```

IV4 *CONFIDENCE INTERVAL OF THE RATIO OF VARIANCES AND TEST FOR HOMOGENEITY OF VARIANCES BETWEEN NORMAL POPULATIONS*

Description

It calculates the confidence interval of the ratio of variances of Normal populations and the test for homogeneity of variances.

Usage

```
IV4(data, variables, varSel, group=NULL, alfa=0.05, file="Output.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Data file.
variables	Variable or variables for which the confidence interval will be calculated for the ratio of variances and the test of homogeneity.
varSel	It specifies the variable or variables that two studied groups have. It is important to mention that if the variable was previously defined as grouping variable, it can not be used in this section. If the variable has more than two groups, it is necessary to specify which are the groups that are going to be studied.
group	Variables that gather data for calculations. In case of selecting NULL, there would be no grouping and it would be calculated considering all the data of the selected variables.
alfa	Confidence level.
file	CSV FILE. Output file name.
na	CSV FILE. Text used in cells without data.

`dec`	CSV FILE. Defines whether the comma "," or dot "." is used as decimal separator.
`row.names`	CSV FILE. Logical value that specifies whether identifiers are placed in rows or a vector with a text for each of the rows.

Details

IV. CONFIDENCE INTERVALS

IV.4. CONFIDENCE INTERVAL OF THE RATIOS OF VARIANCES OF TWO NORMAL POPULATIONS

To determine the confidence interval of the ratio of variances is important, for example when we want to compare the variability of two methods at the time of making a biochemical analysis or the variability of two devices that give us the same type of measure. The comparison of variances between the variables is also important because one of the requirements of the parametric tests of contrasts of homogeneity, *t* test (IX1,IX2), ANOVA (IX3), ANCOVA (IX4), etc., is the fact that there is homogeneity of variances between the variables that are compared.

The interval of the ratio of variances at the level of confidence $1 - \alpha \left(I^{1-\alpha}_{\sigma_1^2/\sigma_2^2} \right)$ is calculated as follows:

$$I^{1-\alpha}_{\sigma_1^2/\sigma_2^2} = \left(\frac{s_1^2}{s_2^2} \frac{1}{F_{\frac{\alpha}{2};n_1-1,n_2-1}}, \frac{s_1^2}{s_2^2} F_{\frac{\alpha}{2};n_2-1,n_1-1} \right)$$

A confidence interval for the standard deviations at the level of confidence $1 - \alpha \left(I^{1-\alpha}_{\sigma_1/\sigma_2} \right)$ is calculated as follows:

$$I^{1-\alpha}_{\sigma_1/\sigma_2} = \left(\frac{s_1}{s_2} \frac{1}{\sqrt{F_{\frac{\alpha}{2};n_1-1,n_2-1}}}, \frac{s_1}{s_2} \sqrt{F_{\frac{\alpha}{2};n_2-1,n_1-1}} \right)$$

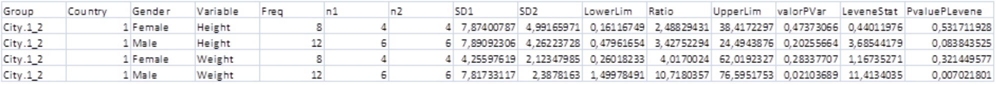
The critical values of the distribution *F* are used (Table 4 of Appendix I Guisande et al., 2011).

EXAMPLE

Data of height and weight in men and women from different countries and cities. The objective is to estimate the interval of confidence of the ratio of variances in weight and height, in men and women, between different cities, and determine if the variances are homogeneous.

For example, in the case of comparing the height of men between the cities 1 and 2, the confidence interval for the ratio of variances is between 0.479 and 24.494. Therefore, the probability for the ratio of variances (p = 0.202), as the Levene test (p = 0.083) and Brown and Forsythe's test (p = 0.217) are greater than 0.05, so it is accepted the null hypothesis that there are no significant differences in the variances of the height of men between the cities 1 and 2, i.e., there is homogeneity of variances.

However, when comparing the weight of men between the cities 1 and 2, both the probability for the ratio of variances (p = 0.021) and of the Levene test (p = 0.007) are less than 0.05, but not Brown & Forsythe's test (p = 0.112), therefore, it is reasonable to reject the null hypothesis and, therefore, that there are significant differences between the variances of the height of men between the cities 1 and 2, i.e., there is no homogeneity of variances.



Group	Country	Gender	Variable	Freq	n1	n2	SD1	SD2	LowerLim	Ratio	UpperLim	valorPVar	LeveneStat	PvaluePLevene
City.1_2	1	Female	Height	8	4	4	7,87400787	4,99165971	0,16116749	2,48829431	38,4172297	0,47373066	0,44011976	0,531711928
City.1_2	1	Male	Height	12	6	6	7,89092306	4,26223728	0,47961654	3,42752294	24,4943876	0,20255664	3,68544179	0,083843525
City.1_2	1	Female	Weight	8	4	4	4,25597619	2,12347985	0,26018233	4,0170024	62,0192327	0,28337707	1,16735271	0,321449577
City.1_2	1	Male	Weight	12	6	6	7,81733117	2,3878163	1,49978491	10,7180357	76,5951753	0,02103689	11,4134035	0,007021801

Value

A CSV file is obtained with a confidence interval for the ratio of variances and the Levene test and Brown & Forsythe test, which allows to determine if there is a homogeneity of variances in the analyzed groups.

Examples

```
## Not run:

data(ZIV1)

IV4(data=ZIV1, variables=c("Height","Weight"), group=c("Country","Gender"),
alfa=0.05, varSel=list(c("City","1","2"),c("City","1","3")))


## End(Not run)
```

IX1 t-*TEST FOR INDEPENDENT SAMPLES*

Description

It applies the test of the *t* for independent samples and data can be shown by using a boxplot or a beanplot.

Usage

```
IX1(data, variables, factor, pop1, pop2, trans=NULL, graph="Boxplot",
PAR=NULL, ResetPAR=TRUE, YLAB=NULL, XLAB=NULL, OrderCat=NULL, LabelCat=NULL,
COLOR=NULL, BOXPLOT=NULL, BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL,
MTEXT= NULL, TEXT=NULL, file="Output.txt")
```

Arguments

`data`	Data file.
`variables`	Dependent variable.
`factor`	It defines which is the variable that acts as a factor.
`pop1`	First group of variable population to be compared.
`pop2`	Second group of vaiable population to be compared.

`trans` Type of transformation that is applied to the data:
1. NULL (untransformed)
2. 1/x2
3. 1/x
4. LN
5. LOG
6. SQR (square root)
7. x2
8. x3
9. EXP (exponential)
10. ASN (arcsine)

`graph` If it is NULL, there is no graph and the other options are "Boxplot" or "Beanplot".

`PAR` Accessing the function PAR which allows to modify many different aspects of the chart.

`ResetPAR` If FALSE, the conditions are not placed by default in the function PAR and those defined by the user on previous graphics are maintained.

`YLAB` A legend for the Y axis in the boxplot and beanplot.

`XLAB` A legend for the X axis in the boxplot and beanplot.

`OrderCat` It allows to specify a vector with the order in which categories are displayed in the graph.

`LabelCat` It allows to specify a vector with the names of the categories of the graph.

`COLOR` Vector with the color of the categories of the chart.

`BOXPLOT` It allows to specify the characteristics of the boxplot.

`BEANPLOT` It allows to specify the characteristics of the beanplot.

`LEGEND` It allows to include a legend.

`AXIS` It allows to add axes.

`MTEXT` It allows to add texts in the margins of the chart.

`TEXT` It allows to add a text in any area of the inner part of the chart.

`file` TXT FILE. Name of the output file with the results.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

For the contrast of homogeneity, in the case of quantitative variables, there are two types of analysis, the **parametrics**, which come from a model that requires them to meet certain assumptions about the parameters and the probability distribution of the population from which the sample was drawn, and the **non-parametrics**, which are less restrictive to be applied. It should be borne in mind that the parametric statistics is more accurate, but the chance of being applied, as discussed below, are limited.

IX.1 PARAMETRIC TEST

IX.1.1. Requirements

Before using parametric tests, it is essential to check certain conditions (assumptions) for its application. The most important to consider are:

1. Normal Distribution of the populations.

2. Equality of variances.

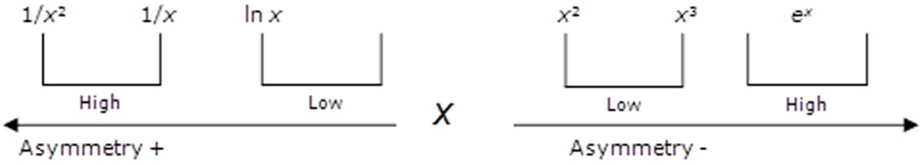
For the verification of the hypothesis of normality, it is necessary to perform a goodness of fit test and the Kolmogorov-Smirnov or the Shapiro-Wilk test, explained in the VI3 function, can be used. The normality test should be done on the residuals or differences between the observed values of the dependent variable and the values that predicts the model once estimated parameters. If the residuals are Normal, also the dependent variable will have Normal distribution at each level of each factor. The assumption of equal variances is verified commonly as detailed in the IV4 function or as explained in the examples discussed below.

If our data do not meet the assumption of Normal distribution, the variable would be transformed to assume a Normal distribution as discussed in the following section. In general, the transformation that "normalizes" the data also achieves equal variances. In the event that this fails to reach the Normal distribution and/or the homogeneity of variances, even after transforming the data, the best alternative would be to use a non-parametric method, although sometimes some parametric tests apply not assuming equal variances, as for example the t-test as modified by Welch-Satterthwaite.

IX.1.2. Transformations

If the variable is transformed, there will be various possibilities depending on the type of distribution (asymmetric positive or negative, see function II1). The literature speaks of the so-called ladder of the transformations of Tukey, which shows the type of transformation recommended depending on the intensity of the asymmetry or the direction in which extreme cases go (Sanchez, 1999). Figure IX.1 is a modification of the chart made by Erickson & Nosanchuk (1977) and adapted by Guisande et al. (2011).

Figure IX.1. Ladder of transformations.



To choose the appropriate transformation, the asymmetry and kurtosis should be taken into account. As the chart shows only the asymmetry, the influence of kurtosis may give rise to values lower than the average require a square root transformation and values higher than the average require squaring. The transformation that makes closer to zero simultaneously both coefficients must be chosen.

IX.1.3. *t*-test

IX.1.3.1. Independent samples

The *t*-test is the most common method to evaluate the differences between the means of two independent groups, for example, two groups of fish subjected to different diets. For this test, the subject should ideally be allocated randomly to two groups, so that any difference in response is due to the treatment (or lack of treatment) and not to other factors. To apply this analysis, it must be ensured that differences in other factors are not masking or enhancing a significant difference between the means. The null hypothesis (H_0) which is commonly used is that the means are equal. In general it is not required that the number of observations in both groups is equal. If the H_0 is certain and equal variances are assumed, the statistical (*t*-test) will follow a *t* distribution of Student with n_x + n_y - 2 degrees of freedom. There is also the option of using the statistical *t* of Student not assuming equal variances, when working with the degrees of freedom of Welch-Sattherthwaite .

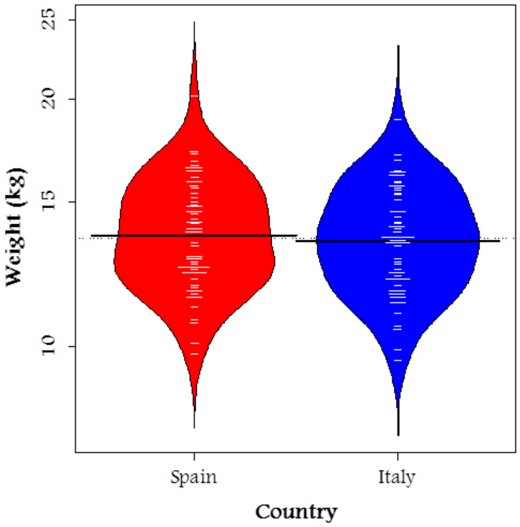
FUNCTIONS

For the beanplot, the beanplot function of the beanplot package is used (Kampstra, 2008; Kampstra, 2013). The shapiro.test function of the base package stats to perform the Shapiro-Wilk test. Levene's test is performed with the levene.test function of the lawstat package(Gastwirth et al., 2013). The asymmetry and kurtosis are performed with the skewness and kurtosis functions, respectively, of the package e1071 (Meyer et al., 2014).

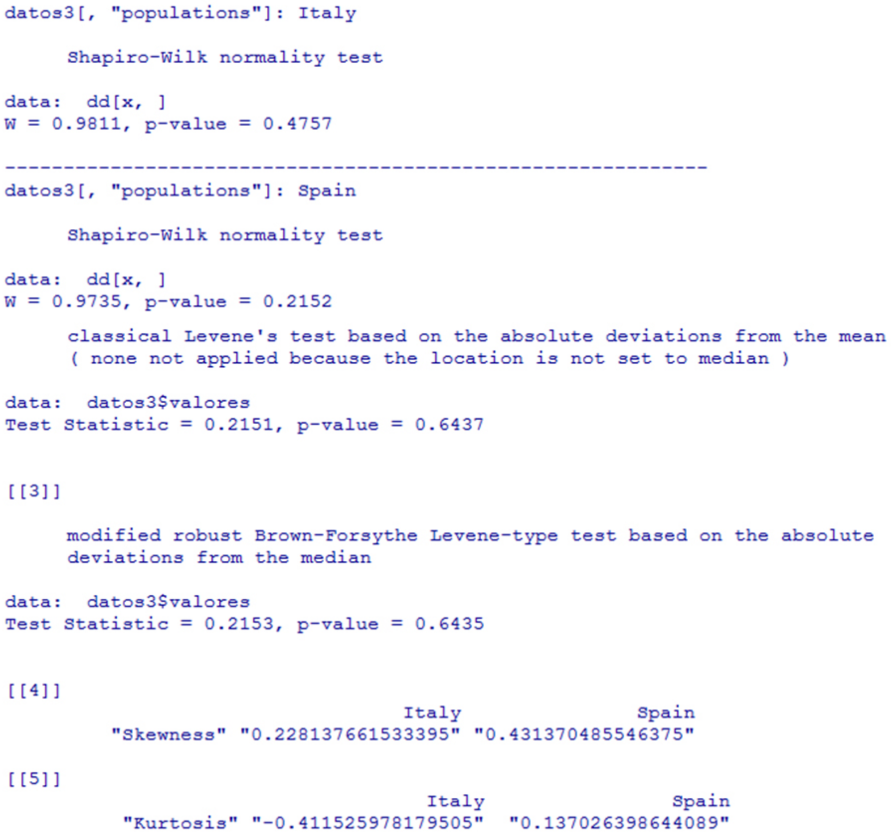
EXAMPLE

Data on weight and height of boys and girls between 2 and 4 years in Italy and Spain. The aim of this work is to determine if there are significant differences in weight, jointly considering boys and girls in both countries. Figure IX.2 shows the median and distribution of the weight data of children in Italy and Spain.

Figure IX.2. Beanplot comparing the weight of children between Spain and Italy.

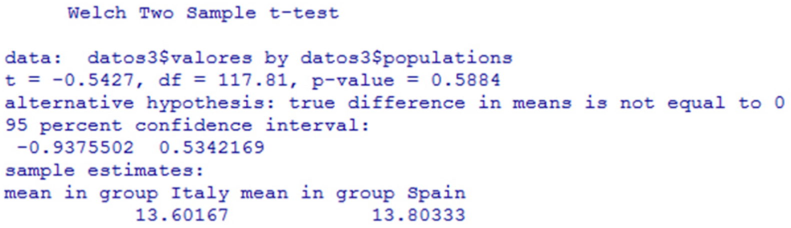


The Shapiro-Wilk test shows that the two populations of children in Spain (p = 0.215) and Italy (p = 0.476) have normal distribution. It will also fulfill the requirement of homogeneity of variances, both considering the average (p = 0.644) as the median (p = 0.643). Therefore, it is not necessary to perform the analysis again transforming the dependent variable. In the case of failure to fulfill any of the two requirements, there is also information on the asymmetry and the kurtosis of each of the populations, which would help to choose the best processing to perform.



```
datos3[, "populations"]: Italy

     Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9811, p-value = 0.4757

------------------------------------------------------------
datos3[, "populations"]: Spain

     Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9735, p-value = 0.2152
     classical Levene's test based on the absolute deviations from the mean
     ( none not applied because the location is not set to median )

data:  datos3$valores
Test Statistic = 0.2151, p-value = 0.6437


[[3]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$valores
Test Statistic = 0.2153, p-value = 0.6435


[[4]]
                                 Italy                Spain
         "Skewness" "0.228137661533395" "0.431370485546375"

[[5]]
                                   Italy                 Spain
          "Kurtosis" "-0.411525978179505"  "0.137026398644089"
```

Finally, the test of the *t* shows no significant difference in the weight of the boys and girls between Spain and Italy (t =-0.54, df = 117.8, p = 0.588), with mean values of 13.8 and 13.6 kg, respectively.



```
     Welch Two Sample t-test

data:  datos3$valores by datos3$populations
t = -0.5427, df = 117.81, p-value = 0.5884
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9375502  0.5342169
sample estimates:
mean in group Italy mean in group Spain
           13.60167            13.80333
```

Value

A TXT file is obtained with the results of the *t*-test and it can render a boxplot or a beanplot.

References

Erikson, F.H. & Nosanchuk, T.A. (1977) *Understandig data*. McGraw Hill. Toronto.

Gastwirth, J.L., Gel, Y.R., Hui, W.L.W., Lyubchich, V., Miao, W. & Noguchi, K. (2013). An R package for biostatistics, public policy, and law. R package version 2.4.1. Available at: http://CRAN.R-project.org/package=lawstat.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Kampstra, P (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. & Lin, C.C (2014) Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. Available at: http://CRAN.R-project.org/package=e1071.

Sánchez, J.J. (1999) *Manual de análisis estadístico de los datos*. Alianza Editorial. Madrid.

Examples

```
## Not run:

data(ZIX1)

#Comparison of the weight of children between Italy and Spain

IX1(data = ZIX1, variables = "Weight", factor = "Country", pop1 = "Spain",
pop2 = "Italy", graph = "Beanplot", PAR = c("cex.lab=1.6", "font.lab=2",
"cex.axis=1.4","family = 'Andalus'", "omi=c(0,0.7,0,0)"),
BEANPLOT = c("col = list(c('red','white'), c('blue','white'))",
"ll = 0.04", "ylab = 'Weight (kg)'", "xlab='Country'"))


## End(Not run)
```

IX10 *CONTRASTS OF NON-PARAMETRIC HOMOGENEITY FOR K INDEPENDENT SAMPLES*

Description

The sum of ranks of Kruskal-Wallis test is applied. In addition, the data can be represented with a boxplot or a beanplot.

Usage

```
IX10(data, variables, Factor, ResetPAR=TRUE, graph="Boxplot", PAR=NULL, YLAB=NULL,
XLAB=NULL, OrderCat=NULL, LabelCat=NULL, COLOR=NULL, BOXPLOT=NULL,
BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL, file="Output.txt")
```

Arguments

`data`	Data file.
`variables`	Dependent variable.
`Factor`	It defines which is the variable that acts as a factor.
`ResetPAR`	If FALSE, the default conditions of the function PAR are not placed and are those defined by the user in previous graphs.
`graph`	If NULL, there is no graph and the other options are "Boxplot" or "Beanplot".
`PAR`	It accesses the function PAR which allows to modify many different aspects of the graph.
`YLAB`	Legend of the Y axis in the Boxplot and beanplot.
`XLAB`	Legend of the X axis in the Boxplot and beanplot.
`OrderCat`	It allows to specify a vector with the order in which categories are displayed in the graph.
`LabelCat`	It allows to specify a vector with the names of the categories of the graph.
`COLOR`	Vector with the color of the categories of the graph.
`BOXPLOT`	It allows to specify the characteristics of the boxplot.
`BEANPLOT`	It allows to specify the characteristics of the beanplot.
`LEGEND`	It allows to include a legend.
`AXIS`	It allows to add axes.
`MTEXT`	It allows to add text in the margins of the graph.
`TEXT`	It allows to add text in any area of the inner part of the graph.
`file`	TXT FILE. Output file name with the results.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.2. NON-PARAMETRIC TESTS

IX.2.2. Contrasts for *K*-Independent samples

The most used contrasts are the median and the Kruskal-Wallis test, which are briefly explained below.

IX.2.2.1. Median contrast

This is to determine the median value of our observations including all the k samples to analyze and characterize all the values as positive, if they are above the median and negative if they are below. If the samples were homogeneous, in each one of them, about half of the values should be above the median and the other half below. This will be the null hypothesis (that all the samples come from the same population).

A contingency table 2xk with the number of positive and negative data from each sample is made. A contrast χ^2 of homogeneity is applied with the null hypothesis indicated, i.e., in each sample the expected frequency of positive is equal to the expected frequency of negative.

This contrast is not very sensitive, so it is more accurate using the Kruskal-Wallis test. However, in cases where the orders have artificial limits this contrast, in theory, this is the most suitable.

IX.2.2.2. ANOVA of Kruskal-Wallis

The contrast is analogous to the non-parametric analysis of variance, this is why it is called ANOVA, although the statistic has Chi-square distribution. It is the most widely used test for more than two independent samples, being much more sensitive than the test of the median. On a theoretical level, it is an extension of the U Mann-Whitney contrast for more than two samples. This means that it is a test that measures the central tendency of the samples, having as null hypothesis that the tested populations have the same median.

The complete development of the calculation of this statistic is at Sokal & Rohlf (1981), in which the steps are:

1. Sort the data in ranges (all samples combined).

2. Add the ranges of each sample and calculate the statistic of H that compares with a tabulated value of contrast.

3. If H is greater than the tabulated value the null hypothesis is rejected, indicating that the samples are different.

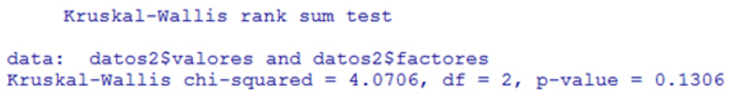
FUNCTIONS

For the graph beanplot, the function beanplot of the beanplot package (Kampstra, 2008; Kampstra, 2013) is used. The Kruskal-Wallis ANOVA is performed with the function kruskal.test of the stats package.

EXAMPLE

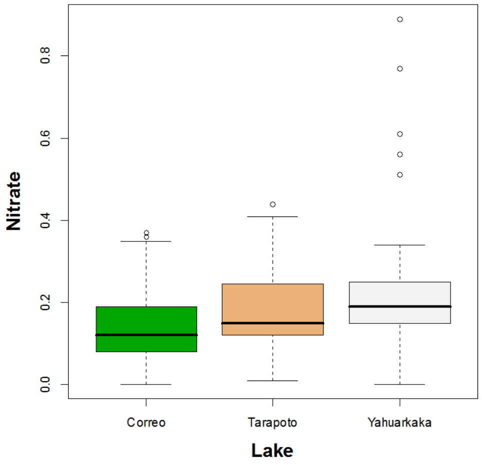
Data from the concentration of nitrite, nitrate and ammonium (μML) in the lakes of the Colombian Amazon in two different months. The aim is to determine whether there are differences in the concentration of nitrate among all lakes.

The test result has a p = 0.131, so the null hypothesis of homogeneity (Figure IX.13) is accepted.



```
        Kruskal-Wallis rank sum test

data:  datos2$valores and datos2$factores
Kruskal-Wallis chi-squared = 4.0706, df = 2, p-value = 0.1306
```

Figure IX.13. Nitrate concentration (μML) in several lakes of the Amazon



Value

A TXT file is obtained with the results of the test and a boxplot or beanplot can be displayed.

References

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Sokal, R.R. & Rohlf, F.J. (1981) *Biometry*. WH Freeman and Company, New York.

Examples

```
## Not run:

data(ZIX7)

IX10(data=ZIX7, variables="Nitrate", Factor="Lake")


## End(Not run)
```

IX11 *CONTRASTS OF HOMOGENEITY FOR TWO NON-PARAMETRIC DEPENDENT SAMPLES*

Description

The Wilcoxon test is applied for related pairs. In addition, the data can be represented with a boxplot or a beanplot.

Usage

```
IX11(data, varTime, graph="Boxplot", PAR=NULL, ResetPAR=TRUE, YLAB="Dependent variable",
XLAB="Time", LabelCat=NULL, COLOR=NULL, BOXPLOT=NULL, BEANPLOT=NULL,
LEGEND=NULL, AXIS=NULL, MTEXT=NULL, TEXT=NULL, file="Output.txt")
```

Arguments

`data`	Data File.
`varTime`	Columns in which the dependent variable has been measured over time.
`graph`	If NULL, there is no graph and the other options are "Boxplot" or "Beanplot".
`PAR`	It accesses the function PAR which allows to modify many different aspects of the graph.
`ResetPAR`	If FALSE, the default conditions of the function PAR are not placed and are those defined by the user in previous graphs.

`YLAB`	Legend of the Y axis in the boxplot and beanplot.
`XLAB`	Legend of the X axis in the boxplot and beanplot.
`LabelCat`	It allows to specify a vector with the names of the categories of the graph.
`COLOR`	Vector with the color of the categories of the graph.
`BOXPLOT`	It allows to specify the characteristics of the boxplot.
`BEANPLOT`	It allows to specify the characteristics of the beanplot.
`LEGEND`	It allows to include a legend.
`AXIS`	It allows to add axes.
`MTEXT`	It allows to add text in the margins of the graph.
`TEXT`	It allows to add text in any area of the inner part of the graph.
`file`	TXT FILE. Output file name with the results.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.2. NON-PARAMETRIC TESTS

IX.2.3. Contrasts for two dependent samples

The tests most commonly used are the contrast of the signs, and the Wilcoxon test for related pairs.

IX.2.3.1. Contrast of the signs

This contrast measures the direction of the differences between two paired samples, which is very useful when working with paired samples to which an increase or decrease in the measurement can be identified, but it is difficult to quantify this change. In other conditions, the Wilcoxon test is more sensitive.

The number of times that the differences of the values of the two samples are positive, negative or no difference is studied. With these premises the null hypothesis that the samples arehomogeneous can be contrasted, since in that case there will be approximately as many positive as negative differences.

The full development of the test appears in Siegel & Castellan (1988), but the brief steps are:

1. Determining differences (+ or -).

2. Adjustment of the differences observed to a binomial distribution. When the number of data is large (greater than 30), an approximation to the Normal distribution of the binomial distribution (statistical calculation of Z) is used.

3. Determination of the probability associated with the statistic of contrast to accept or reject the null hypothesis that the samples are homogeneous.

IX.2.3.2. Wilcoxon test for related pairs

The test is similar to the t of Student for related samples, being almost as powerful as this. In regard to the sensitivity, it is a contrast much more powerful than the previous one because, although it also uses the differences between the values for each case, the absolute value of the differences are sorted in ranges (not just positive and negative), having more information about the differences in the previous case.

If samples are homogeneous, the null hypothesis, the sum of the ranges of the positive differences must be similar to the sum of the ranges with a negative value (randomness of the differences).

A detailed description of this test is at Siegel & Castellan (1988), the main steps are:

1. Sort the values of the samples in ascending order and sort in ranges the difference in absolute value between both samples.

2. Calculate a test statistic T and T'.

3. If T or T' is less than or equal to the threshold quantities listed in the so-called Table of Wilcoxon, the null hypothesis that the variables are homogeneous is rejected.

In the event that the sample size is high (greater than 100), an approximation to the Normal of the statistical T can be performed. The value Z of the Normal distribution that returns the probability of the contrast can be calculated.

FUNCTIONS

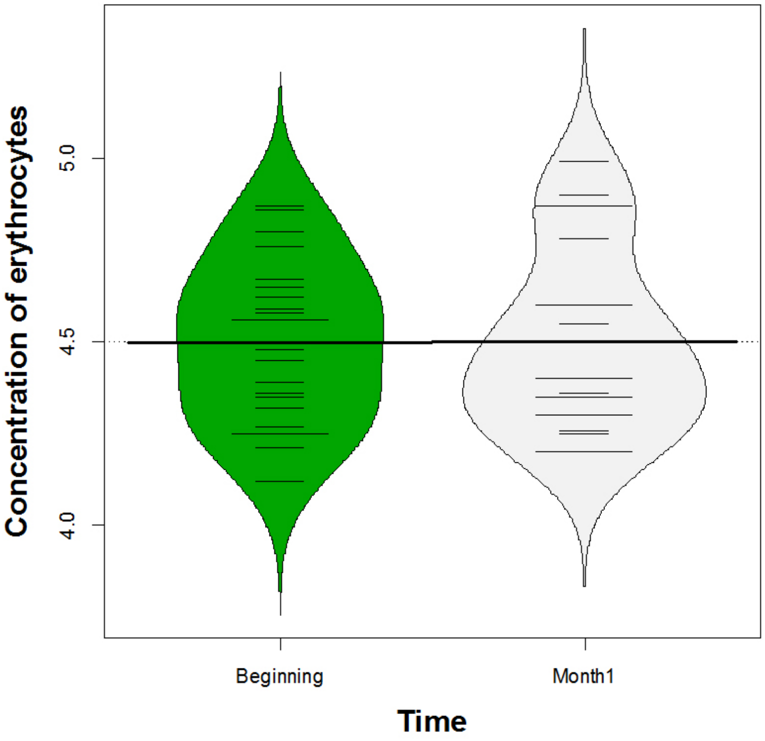
For the beanplot, the function beanplot of the package beanplot (Kampstra, 2008; Kampstra, 2013) is used. The Wilcoxon test is done with the function wilcox.test of the stats package.

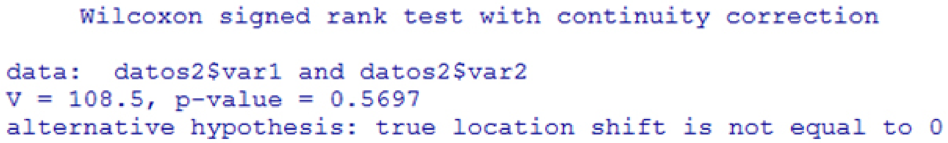
EXAMPLE

Concentration data of erythrocytes (in millions per cubic millimeter), of several men and women who were undergoing a treatment to increase the concentration of red blood cells. The objective is to determine if the concentration is different between the beginning and at the end of a month.

Figure IX.14 shows that the concentration of erythrocytes is very similar to the beginning and at the end of a month, as well as demonstrates the Wilcoxon test (p = 0.57).

Figure IX.14. Erythrocyte concentration (in millions per cubic millimeter) in men and women at the beginning and after a month of treatment.





```
        Wilcoxon signed rank test with continuity correction

data:  datos2$var1 and datos2$var2
V = 108.5, p-value = 0.5697
alternative hypothesis: true location shift is not equal to 0
```

Value

A TXT file is obtained with the results of the test of Wilcoxon for two dependent samples and a boxplot or beanplot can be displayed.

References

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Siegel, S. & Castellan, N.J. Jr. (1988) *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Examples

```
## Not run:

data(ZIX9)

#Variation in the concentration of erythrocytes over time

IX11(data=ZIX9, varTime=c("Beginning", "Month1"),graph = "Beanplot",
XLAB="Time", YLAB="Concentration of erythrocytes")

## End(Not run)
```

IX12 *CONTRASTS OF HOMOGENEITY FOR NON-PARAMETRIC K INDEPENDENT SAMPLES*

Description

Friedman ANOVA is applied to k-related samples. In addition, the data can be represented with a boxplot or a beanplot.

Usage

```
IX12(data, varTime, Case, graph="Boxplot", PAR=NULL, ResetPAR=TRUE,
YLAB="Dependent variable", XLAB="Time", LabelCat=NULL, COLOR=NULL, BOXPLOT=NULL,
BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT=NULL, TEXT=NULL, file="Output.txt")
```

Arguments

`data`	Data file.
`varTime`	Columns in which the dependent variable has been measured over time.
`Case`	Variable that identifies each case.
`graph`	If NULL, there is no graph and the other options are "Boxplot" or "Beanplot".
`PAR`	It accesses the function PAR which allows to modify many different aspects of the graph.
`ResetPAR`	If FALSE, the default conditions of the function PAR are not placed and are those defined by the user in previous graphs.
`YLAB`	Legend of the Y axis in the boxplot and beanplot.
`XLAB`	Legend of the X axis in the boxplot and beanplot.
`LabelCat`	It allows to specify a vector with the names of the categories of the graph.
`COLOR`	Vector with the color of the categories of the graph.
`BOXPLOT`	It allows to specify the characteristics of the boxplot.
`BEANPLOT`	It allows to specify the characteristics of the beanplot.
`LEGEND`	It allows to include a legend.
`AXIS`	It allows to add axes.
`MTEXT`	It allows to add text in the margins of the graph.
`TEXT`	It allows to add text in any area of the inner part of the graph.
`file`	TXT FILE. Output file name with the results.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.2. NON-PARAMETRIC TESTS

IX.2.4. Contrasts for *K*-dependent samples

The most commonly used test is the Friedman ANOVA. This test contrasts the hypothesis that measures come from the same populations, when data from each of the samples have been transformed into ranks in ascending order.

This test is analogous to ANOVA for repeated samples so, even thought it is a contrast χ^2, it is called ANOVA.

The steps of this test are described by Siegel & Castellan (1988). Briefly they are:

1. Assign ranges within each sample case.

2. Add the ranges of each variable for the entire sample.

3. Calculate a test statistic χ^2 and compare it with the tabulated value χ^2 critical with a - 1 degrees of freedom, being a the number of measured values for each sample.

FUNCTIONS

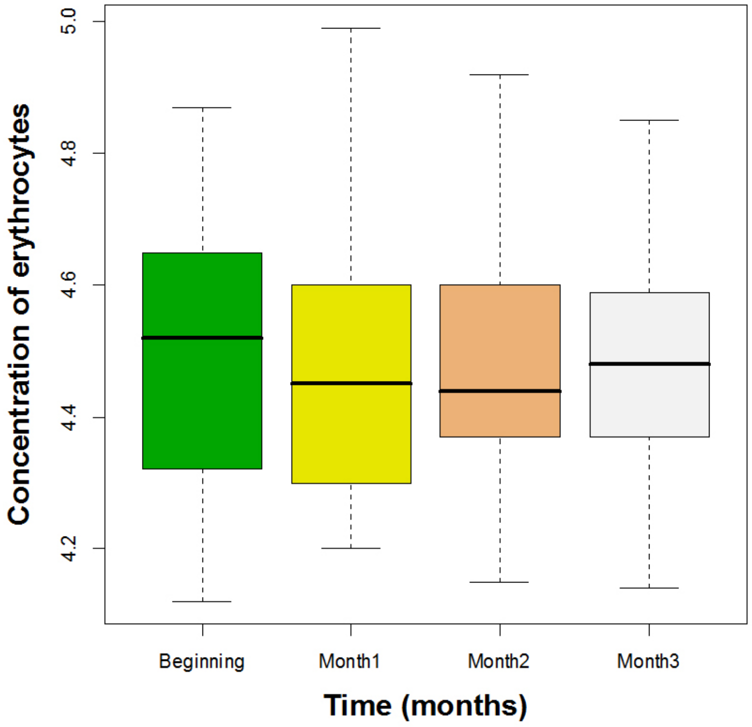
For the beanplot, the function beanplot of the package beanplot (Kampstra, 2008; Kampstra, 2013) is used. The ANOVA of Friedman is performed with the function friedman.test of the stats package.

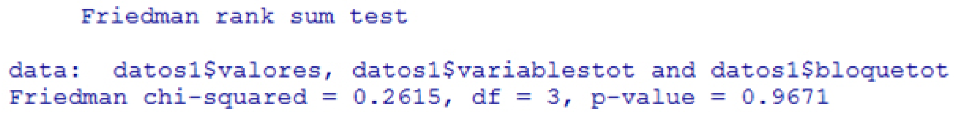
EXAMPLE

Concentration data of erythrocytes (in millions per cubic millimeter), several men and women who were undergoing a treatment to increase the concentration of red blood cells. The objective is to determine if the concentration of red blood cells is different considering all times analyzed, and considering the set of men and women.

Figure IX.15 shows that the concentration of erythrocytes is very similar at the beginning and in the following months and indeed the results of the contrast of Friedman show that, in the set of men and women, the concentration of red blood cells does not change over time (p = 0.967).

Figure IX.15. Erythrocyte concentration (in millions per cubic millimeter) in men and women at the beginning and in the next three months.





```
        Friedman rank sum test

data:  datos1$valores, datos1$variablestot and datos1$bloquetot
Friedman chi-squared = 0.2615, df = 3, p-value = 0.9671
```

Value

A TXT file is obtained with the results of the ANOVA of Friedman and a boxplot or beanplot can be displayed.

References

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Siegel, S. & Castellan, N.J. Jr. (1988) *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Examples

```
## Not run:

data(ZIX10)

#Variation in the concentration of erythrocytes over time

IX12(data=ZIX10, varTime=c("Beginning", "Month1", "Month2", "Month3"),
Case="Individual", XLAB="Time (months)", YLAB="Concentration of erythrocytes")

## End(Not run)
```

IX13 *ANALYSIS OF COVARIANCE (ANCOVA) NON-PARAMETRIC*

Description

An analysis of covariance using a statistical ANOVA type (Wang & Ye, 2010; Wang, 2013) is applied. In addition, the relationship between the dependent variable and the covariate can be plotted, differentiating between the groups of the factor.

Usage

```
IX13(data, variables, Factor, Covariable, T.AOV=NULL, ResetPAR=TRUE,
graph=TRUE, PAR=NULL, YLAB=NULL, XLAB=NULL, MAIN=NULL, CEXPCH=1.3,
COLOR=NULL, PCH=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL,
file="Output.txt")
```

Arguments

`data` Data file.

`variables` Dependent variable.

`Factor` Factor.

`Covariable` Covariate.

`T.AOV` It accesses the funtion T.aov.

`ResetPAR` If FALSE, the conditions are not placed by default in the PAR function and are those defined by the user in previous graphics.

`graph` If NULL, the graph is not performed.

`PAR` It accesses the function PAR that allows to modify many different aspects of the graph.

`YLAB` Legend of the Y axis.

`XLAB` Legend of the X axis.

`MAIN` Graph title.

`CEXPCH` Size of the graph symbols.

`COLOR` This allows to change the colors of the graph, but it must be as many different groups as the factor has.

`PCH` Vector with the symbols on the graph, but they must be as many as different groups have the factor. If it is NULL, they are automatically calculated starting with the symbol 15.

`LEGEND` It allows to include a legend to a graph.

`AXIS` It allows to add axis to a graph.

`MTEXT` It allows to add text in the margins of the graph.

`TEXT` It allows to add text in any area of the inner part of the chart.

`file` TXT FILE. Output file name with the results.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.2. NONPARAMETRIC TESTS

IX.2.5. Non-parametric ANCOVA

There are also non-parametric methods to perform an analysis of covariance. The package fANCOVA (Wang & Ye, 2010; Wang, 2013) contains some functions that perform this analysis based on the idea of comparing two curves to check their equality or parallelism, which is valid for regression curves or for density functions, and therefore can be considered equivalent to a parametric analysis of covariance.

FUNCTIONS

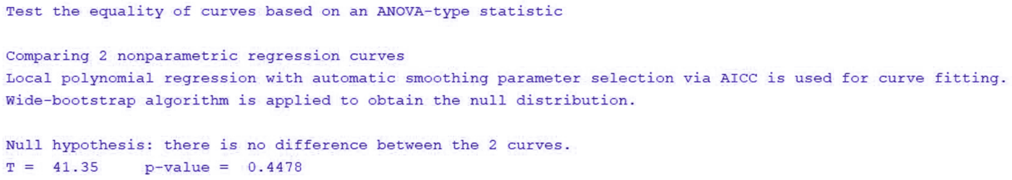
It uses the function T.aov of package fANCOVA (Wang & Ye, 2010; Wang, 2013), which uses a statistical type (ANOVA) to perform the non-parametric ANCOVA.

EXAMPLE 1

It compares the relationship between minimum and maximum temperature between the years 1990 and 2000, considering the daily records of three cities in Spain: Huelva, Palma de Mallorca and Vigo. The objective is to determine if there are differences in the maximum temperature between the two years, once eliminated the possible effect of the minimum temperature that is considered as a covariate.

It is important to mention that the variables must be numeric, i.e., it has been observed that if variables are used with text, the function can result in errors. If it is necessary to make the graph with variables of text(to better fit the legend), the parametric ANCOVA of the function IX4 can be used.

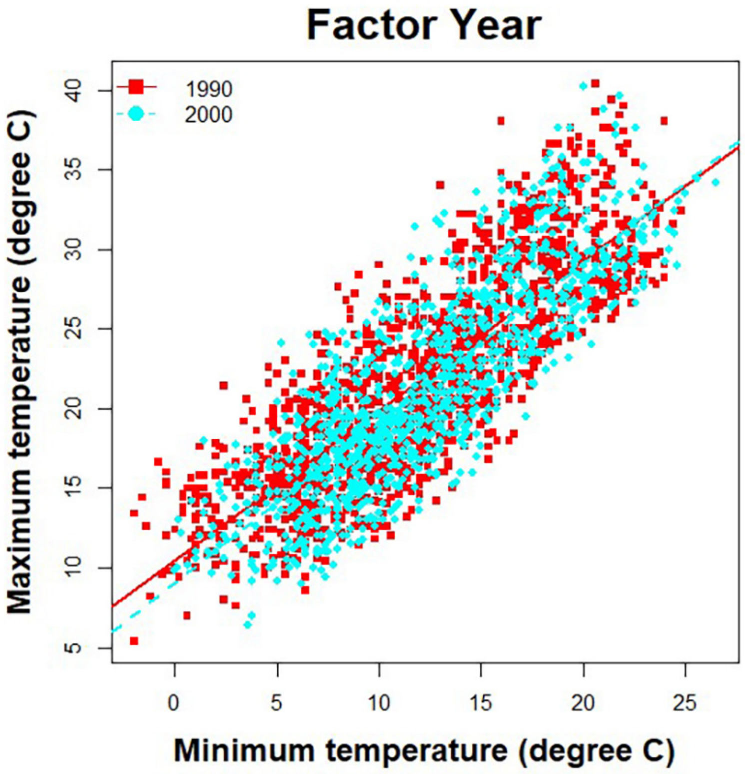
The results show that there were no significant differences in the relationship between minimum and maximum temperature between the years 1990 and 2000 (p = 0.383). It is important to note that if the analysis is repeated, the probability value can be changed, i.e., it does not always give the same result.



```
Test the equality of curves based on an ANOVA-type statistic

Comparing 2 nonparametric regression curves
Local polynomial regression with automatic smoothing parameter selection via AICC is used for curve fitting.
Wide-bootstrap algorithm is applied to obtain the null distribution.

Null hypothesis: there is no difference between the 2 curves.
T =  41.35     p-value =  0.4478
```

Figure IX.16 shows that there is a clear relationship between maximum and minimum temperature considering the three cities.

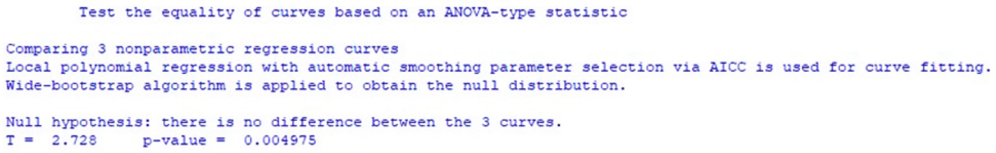
Figure IX.16. Relationship between minimum and maximum temperature in 1990 and 2000, in three cities in Spain: Huelva, Palma de Mallorca and Vigo.



EXAMPLE 2

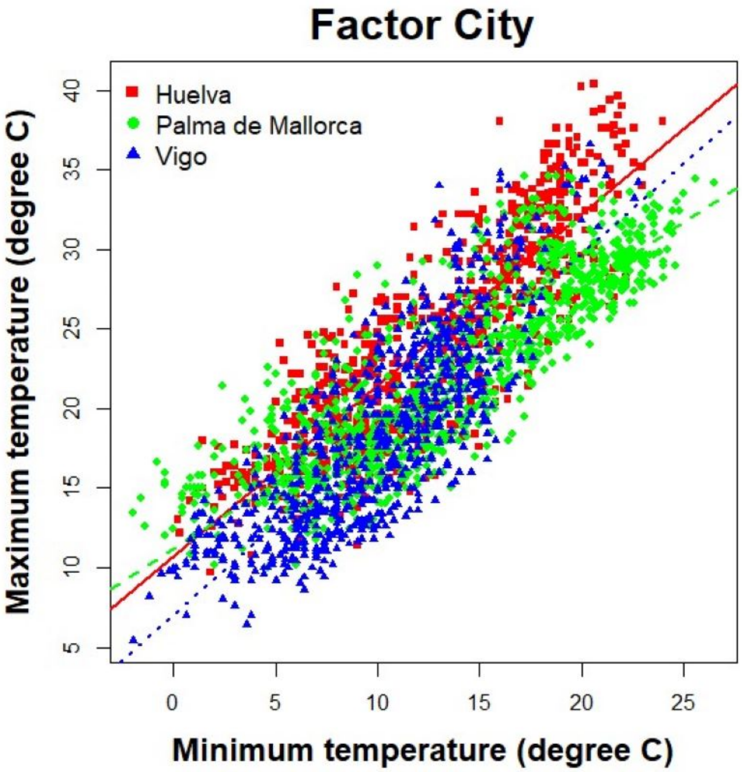
It compares the relationship between minimum and maximum temperature with daily records in 1990 and 2000, between three cities in Spain: Huelva, Palma de Mallorca and Vigo. The objective is to determine if there are differences in temperature between the three cities, removing the possible effect of the minimum temperature which is considered as a covariate.

Figure IX.17 shows that there are clear differences in the relationship between minimum and maximum temperature between the three cities, which is confirmed by contrast, since p = 0.005.



```
        Test the equality of curves based on an ANOVA-type statistic

Comparing 3 nonparametric regression curves
Local polynomial regression with automatic smoothing parameter selection via AICC is used for curve fitting.
Wide-bootstrap algorithm is applied to obtain the null distribution.

Null hypothesis: there is no difference between the 3 curves.
T =  2.728     p-value =  0.004975
```

Figure IX.17. Relationship between the minimum and maximum temperature, considering the years 1990 and 2000, in three cities in Spain: Huelva, Palma de Mallorca and Vigo.



Value

A TXT file is obtained with the results of non-parametric ANCOVA and a graph can be represented linking the dependent variable with the covariate, besides differentiating between factor groups.

References

Wang. X.F. & Ye, D. (2010). On nonparametric comparison of images and regression surfaces. *Journal of Statistical Planning and Inference*, 140; 2875-2884.

Wang, X.F. (2013). Nonparametric Analysis of Covariance. R package version 0.5-1. Available at: http://CRAN.R-project.org/package=fANCOVA.

Examples

```
## Not run:

data(ZIX11)

#EXAMPLE 1
IX13(data=ZIX11, variables="T.Max", Factor="Year", Covariable="T.Min",
XLAB="Minimum temperature (degree C)", YLAB="Maximum temperature (degree C)",
MAIN="Factor Year", CEXPCH=0.8)

#EXAMPLE 2
IX13(data=ZIX11, variables="T.Max", Factor="City", Covariable="T.Min",
XLAB="Minimum temperature (degree C)", YLAB="Maximum temperature (degree C)", MAIN="Factor City",
CEXPCH=0.8, LEGEND=c("x = 'topleft'" , "legend=c('Huelva', 'Palma de Mallorca','Vigo')",
"bty = 'n'", "pch=c(15,16,17)", "col=rainbow(3)", "cex=1.2"))


## End(Not run)
```

IX2 t-*TEST FOR DEPENDENT SAMPLES*

Description

It applies the test of the *t* for dependent samples and data can be shown by using a boxplot or a beanplot.

Usage

```
IX2(data, variables, factor, pop1, pop2, trans=NULL, graph="Boxplot", PAR=NULL,
ResetPAR=TRUE, YLAB=NULL, XLAB=NULL, OrderCat=NULL, LabelCat=NULL, COLOR=NULL,
BOXPLOT=NULL, BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL,
file="Output.txt")
```

Arguments

`data`	Data file.
`variables`	Dependent variable.
`factor`	It defines which is the variable that acts as a factor.
`pop1`	First group of variable population to be compared.
`pop2`	Second group of variable population to be compared.
`trans`	Type of transformation that is applied to the data: 1. NULL (untransformed) 2. 1/x2 3. 1/x 4. LN

5. LOG
6. SQR (square root)
7. x2
8. x3
9. EXP (exponential)
10. ASN (arcsine)

`graph` If it is NULL, there is no graph and the other options are "Boxplot" or "Beanplot".

`PAR` Accessing the function PAR which allows to modify many different aspects of the chart.

`ResetPAR` If FALSE, the conditions are not placed by default in the function PAR and those defined by the user on previous graphics are maintained.

`YLAB` A legend for the Y axis in the boxplot and beanplot.

`XLAB` A legend for the X axis in the boxplot and beanplot.

`OrderCat` It allows to specify a vector with the order in which categories are displayed in the graph.

`LabelCat` It allows to specify a vector with the names of the categories of the graph.

`COLOR` Vector with the color of the categories of the chart.

`BOXPLOT` It allows to specify the characteristics of the Boxplot.

`BEANPLOT` It allows to specify the characteristics of the beanplot.

`LEGEND` It allows to include a legend.

`AXIS` It allows to add axes.

`MTEXT` It allows to add texts in the margins of the chart.

`TEXT` It allows to add a text in any area of the inner part of the chart.

`file` TXT FILE. Name of the output file with the results.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.1.3. *t*-test

IX.1.3.2. Dependent samples

The *t*-test for related samples compares the means of two variables for a single group. It calculates the differences between the values of each variable and verifies if the average difference is significantly different from zero (Sanchez, 1999).

This test assumes that the samples are dependent, paired or related, therefore, a fundamental requirement is to have an equal number of observations in both variables.

This *t*-test for dependent samples requires no assumption about the variances, but requires that the differences in the values of each pair be normally distributed. If it is not, but the sample size is large, the central limit theorem guarantees that the probability distribution of the mean difference be approximately Normal, which allows to use the Normal distribution instead of the *t* distribution. Given that both distributions are virtually identical when the number of degrees of freedom is very large, this means that with large samples, the test can be applied even if the distribution of the variable deviates from normal.

The dependent samples often appear when evaluating the same variable more than once in each subject in the sample (for example in different time ranges). The *t*-test is not focused on the variability that can occur between individuals, but in the observed differences in a same subject from one moment to the next. This test is also applied in case-control studies where each case individually paired with a control.

FUNCTIONS

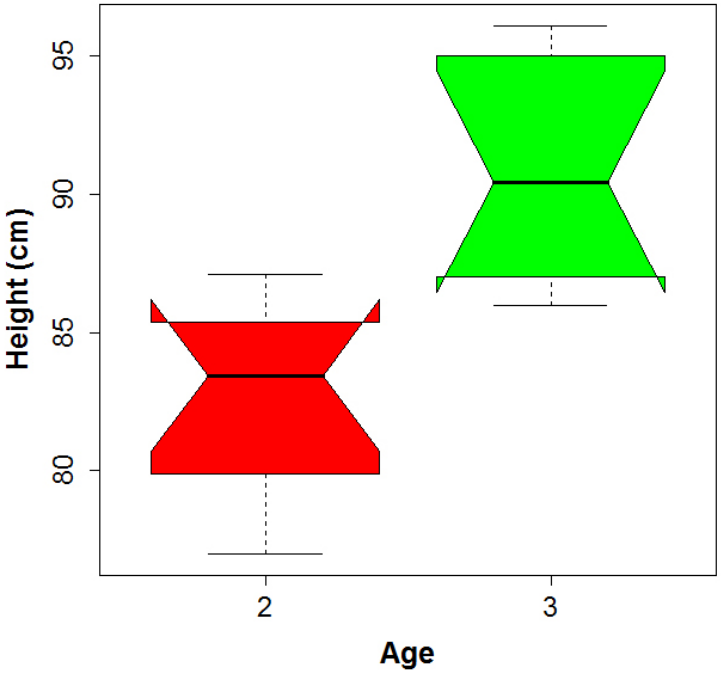
For the beanplot, the beanplot function of the beanplot package is used (Kampstra, 2008; Kampstra, 2013). The shapiro.test function of the base package stats to perform the Shapiro-Wilk test. Levene's test is performed with the levene.test function of the lawstat package(Gastwirth et al., 2013). The asymmetry and kurtosis are performed with the skewness and kurtosis functions, respectively, of the package e1071 (Meyer et al., 2014).

EXAMPLE

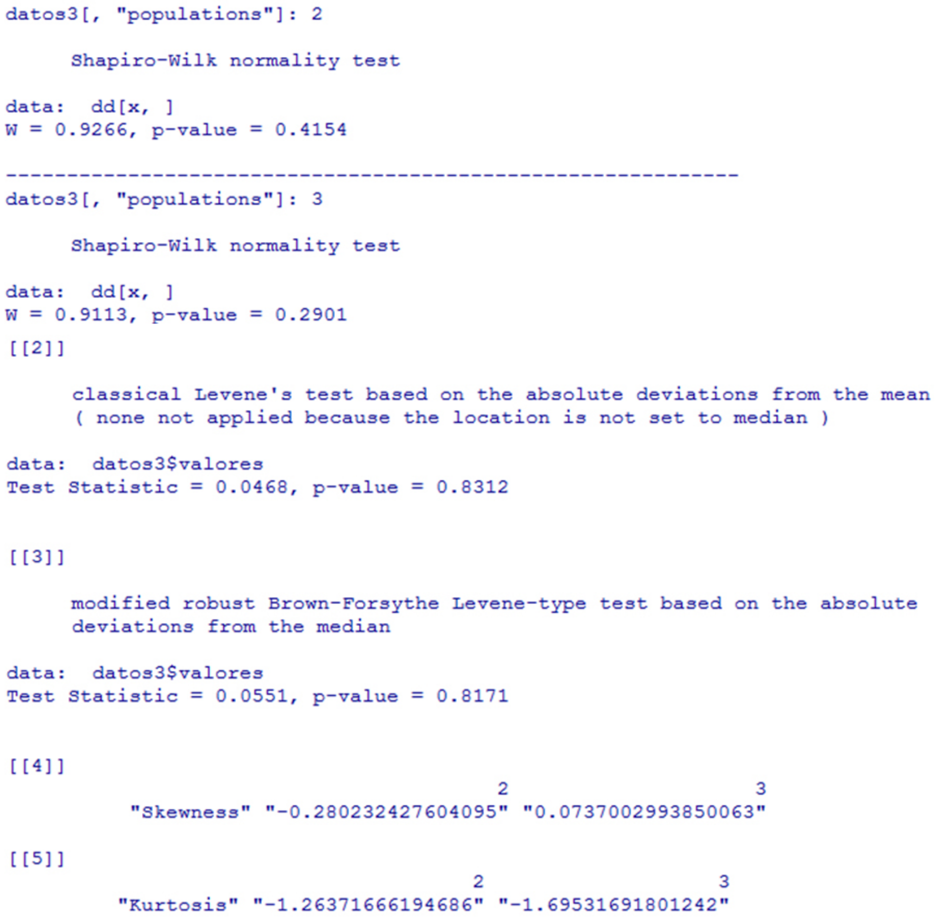
Data on weight and height of boys and girls between 2 and 3 years in Italy and Spain. The objective is to determine if there are significant differences in the height of girls between the ages of 2 and 3 in Italy. Boys and girls that are measured and weighted as they age are always the same.

Figure IX.3 shows clear differences in the height of Italian girls aged 2 and 3 years.

Figure IX.3. Boxplot comparing the height of Italian girls between the ages of two and three.

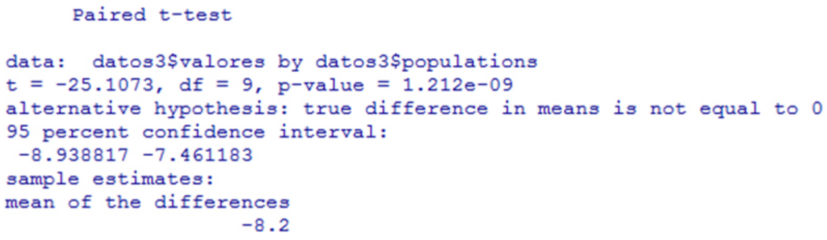


The Shapiro-Wilk test shows that the two populations of girls of 2 (p = 0.415) and 3 (p = 0.29) years have normal distribution. It also fulfills the requirement of homogeneity of variances, considering both the mean (p = 0.831) and median (p = 0.817). Therefore, it is not necessary to perform the analysis again transforming the dependent variable. In the case of failure to fulfill any of the two requirements, there is also information on the asymmetry and the kurtosis of each of the populations, which would help to better choose on the transformation to be performed.



```
datos3[, "populations"]: 2

      Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9266, p-value = 0.4154

------------------------------------------------------------
datos3[, "populations"]: 3

      Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9113, p-value = 0.2901
[[2]]

      classical Levene's test based on the absolute deviations from the mean
      ( none not applied because the location is not set to median )

data:  datos3$valores
Test Statistic = 0.0468, p-value = 0.8312


[[3]]

      modified robust Brown-Forsythe Levene-type test based on the absolute
      deviations from the median

data:  datos3$valores
Test Statistic = 0.0551, p-value = 0.8171


[[4]]
                                       2                    3
          "Skewness" "-0.280232427604095" "0.0737002993850063"

[[5]]
                                     2                   3
         "Kurtosis" "-1.26371666194686" "-1.69531691801242"
```

Finally, the test of the *t* shows that there are significant differences in the height of Italian girls between ages of 2 and 3 (t =-25.1 , df = 9, p < 0.001), with a difference of 8.2 cm.



```
      Paired t-test

data:  datos3$valores by datos3$populations
t = -25.1073, df = 9, p-value = 1.212e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.938817 -7.461183
sample estimates:
mean of the differences
                   -8.2
```

Value

A TXT file is obtained with the results of the *t*-test and it can render a boxplot or a beanplot.

References

Gastwirth, J.L., Gel, Y.R., Hui, W.L.W., Lyubchich, V., Miao, W. & Noguchi, K. (2013). An R package for biostatistics, public policy, and law. R package version 2.4.1. Available at: http://CRAN.R-project.org/package=lawstat.

Kampstra, P (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. & Lin, C.C (2014) Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. Available at: http://CRAN.R-project.org/package=e1071.

Sánchez, J.J. (1999) *Manual de análisis estadístico de los datos*. Alianza Editorial. Madrid.

Examples

```
## Not run:

#Comparison of the height of Italian girls between 2 and 3 years.

data(ZIX1)

data1<-subset(ZIX1, Country=="Italy"  &  Gender=="Girl")

IX2(data=data1, variables="Height", factor="Age", pop1=2, pop2=3,
BOXPLOT= c("xlab='Age'", "ylab='Height (cm)'", "cex.lab=1.6", "col=c('red','green')",
"notch=TRUE", "cex.axis=1.4"), PAR=c("omi=c(0,0.2,0,0)", "font.lab=2"))


## End(Not run)
```

IX3 *ANALYSIS OF VARIANCE(ANOVA)*

Description

An analysis of variance is applied and the data can be displayed with a boxplot or a beanplot.

Usage

```
IX3(data, variables,  Factor1, Factor2=NULL, Factor3=NULL, Factor4=NULL,
SS=3, trans=NULL, ResetPAR=TRUE, mfrow=NULL, graph1="Boxplot",
PAR1=NULL, OrderCat1=NULL, LabelCat1=NULL, COLOR1=NULL, BOXPLOT1=NULL,
BEANPLOT1=NULL, LEGEND1=NULL, AXIS1=NULL, MTEXT1= NULL, TEXT1=NULL,
graph2="Boxplot", PAR2=NULL, OrderCat2=NULL, LabelCat2=NULL, COLOR2=NULL,
BOXPLOT2=NULL, BEANPLOT2=NULL, LEGEND2=NULL, AXIS2=NULL, MTEXT2= NULL,
TEXT2=NULL, graph3="Boxplot", PAR3=NULL, OrderCat3=NULL, LabelCat3=NULL,
COLOR3=NULL,BOXPLOT3=NULL, BEANPLOT3=NULL, LEGEND3=NULL, AXIS3=NULL,
MTEXT3= NULL, TEXT3=NULL, graph4="Boxplot", PAR4=NULL, OrderCat4=NULL,
LabelCat4=NULL, COLOR4=NULL,BOXPLOT4=NULL, BEANPLOT4=NULL, LEGEND4=NULL,
AXIS4=NULL, MTEXT4= NULL, TEXT4=NULL, INTPLOT1.2=TRUE, INTPLOT1.3=TRUE,
INTPLOT1.4=TRUE, INTPLOT2.3=TRUE, INTPLOT2.4=TRUE, INTPLOT3.4=TRUE,
INTERPLOT1.2=NULL, INTERPLOT1.3=NULL, INTERPLOT1.4=NULL, INTERPLOT2.3=NULL,
INTERPLOT2.4=NULL, INTERPLOT3.4=NULL,file="Output.txt")
```

Arguments

`data` Data file.

`variables` Dependent variable.

`Factor1` First factor.

`Factor2` Second factor.

`Factor3` Third factor.

`Factor4` Fourth factor.

`SS` When there are several factors, the decomposition of the sum of squares in parts attributed to each one of the factors is not unique. Four types of sums of squares are commonly used: type I for nested models, type II for balanced models, type III for unbalanced models (also balanced), and type IV for models with empty boxes. The type III is suitable for most applications, so it will be the one used by default. This description is very simplified; interested readers should consult supplementary bibliography for a correct use of the ANOVA models.

`trans` Type of transformation that is applied to the data:

1. NULL (untransformed)
2. 1/x2
3. 1/x
4. LN
5. LOG
6. SQR (square root)
7. x2
8. x3
9. EXP (exponential)
10. ASN (arcsine)

`ResetPAR` If FALSE the conditions are not placed by default in the PAR function and those defined by the user in previous graphics are kept.

`mfrow` If it is NULL and there are various graphics, these appear in separate windows. If the graphics go into panels, this argument is a vector with the format c(nr, nc) indicating the number of figures per row (nr) and column (nc), by first filling the rows.

`graph1` If NULL, the graphic of factor 1 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR1` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 1.

`OrderCat1` It allows to specify a vector with the order in which the categories are shown in graph 1.

`LabelCat1` It allows to specify a vector with the names of the categories of graph 1.

`COLOR1` Vector with the color of the categories of graph 1.

`BOXPLOT1` It allows to specify the characteristics of the boxplot of factor 1.

`BEANPLOT1` It allows to specify the characteristics of the beanplot of factor 1.

`LEGEND1` It allows to include a legend to the graph of factor 1.

`AXIS1` It allows to add axes to the graph of factor 1.

`MTEXT1` It allows to add texts on the margins of the graph of factor 1.

`TEXT1` It allows to add text in any area of the inner part of the graph of factor 1.

`graph2` If NULL, the graphic of factor 2 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR2` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 2.

`OrderCat2` It allows to specify a vector with the order in which the categories are shown in graph 2.

`LabelCat2` It allows to specify a vector with the names of the categories of graph 2.

`COLOR2` Vector with the color of the categories of graph 2.

`BOXPLOT2` It allows to specify the characteristics of the boxplot of factor 2.

`BEANPLOT2` It allows to specify the characteristics of the beanplot of factor 2.

`LEGEND2` It allows to include a legend to the graph of the factor 2.

`AXIS2` It allows to add axes to the graph of the factor 2.

`MTEXT2` It allows to add text in the margins of the graph of the factor 2.

`TEXT2` It allows to add text in any area of the inner part of the graph of the factor 2.

`graph3` If NULL, the graphic of factor 3 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR3` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 3.

`OrderCat3` It allows to specify a vector with the order in which the categories are shown in graph 3.

`LabelCat3` It allows to specify a vector with the names of the categories of graph 3.

`COLOR3` Vector with the color of the categories of graph 3.

`BOXPLOT3` It allows to specify the characteristics of the boxplot of factor 3.

`BEANPLOT3` It allows to specify the characteristics of the beanplot of factor 3.

`LEGEND3` It allows to include a legend to the graph of the factor 3.

`AXIS3` It allows to add axes to the graph of the factor 3.

`MTEXT3` It allows to add text in the margins of the graph of the factor 3.

`TEXT3` It allows to add text in any area of the inner part of the graph of the factor 3.

`graph4` If NULL, graphic factor 4 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR4` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 4.

`OrderCat4` It allows to specify a vector with the order in which the categories are shown in graph 4.

`LabelCat4` It allows to specify a vector with the names of the categories of graph 4.

`COLOR4` Vector with the color of the categories of graph 4.

`BOXPLOT4` It allows to specify the characteristics of the boxplot of factor 4.

`BEANPLOT4` It allows to specify the characteristics of the beanplot of factor 4.

`LEGEND4` It allows to include a legend to the graph of the factor 4.

`AXIS4` It allows to add axes to the graph of the factor 4.

`MTEXT4` It allows to add text in the margins of the graph of the factor 4.

`TEXT4` It allows to add text in any area of the inner part of the graph of the factor 4.

`INTPLOT1.2` If TRUE, it shows the graph of interaction of factor 1 with the factor 2.

`INTPLOT1.3` If TRUE, it shows the graph of interaction of factor 1 with the factor 3.

`INTPLOT1.4` If TRUE, it shows the graph of interaction of factor 1 with the factor 4.

`INTPLOT2.3` If TRUE, it shows the graph of interaction of factor 2 with the factor 3.

`INTPLOT2.4` If TRUE, it shows the graph of interaction of factor 2 with the factor 4.

`INTPLOT3.4` If TRUE, it shows the graph of interaction of factor 3 with the factor 4.

`INTERPLOT1.2` It allows to specify the characteristics of the graph of the interaction of factor 1 with factor 2.

`INTERPLOT1.3` It allows to specify the characteristics of the graph of the interaction of factor 1 with factor 3.

`INTERPLOT1.4` It allows to specify the characteristics of the graph of the interaction of factor 1 with factor 4.

`INTERPLOT2.3` It allows to specify the characteristics of the graph of the interaction of factor 2 with factor 3.

`INTERPLOT2.4` It allows to specify the characteristics of the graph of the interaction of factor 2 with factor 4.

`INTERPLOT3.4` It allows to specify the characteristics of the graph of the interaction of factor 3 with factor 4.

`file` TXT FILE. Name of the output file with the results.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.1.4. Analysis of variance (ANOVA)

In general, when performing a research we wonder if our samples, we consider independent, belong to the same population. The values of samples commonly differ, the problem would be determining if in spite of these differences, populations are equal, and if variations are due to random as a result of the randomness of our sampling.

The most used parametric test to verify if two or more sample means come from the same population is the analysis of variance(ANOVA). The ANOVA is a general method that can be extended to more than two samples and it can be demonstrated that this matches with the *t* of Student if there are only two samples (Salvarrey, 2000; Azzimonti, 2003).

For the application of ANOVA it is necessary to comply with the following assumptions:

· The samples are random and are independent.

· There are no outliers in the dependent variable, which can be detected with the function XV7 of StatR.

· The distribution of the population from which they were extracted is Normal.

· The variances are equal in each group or level of the factors considered.

The ANOVA calculates the variation of means and estimates the "natural" population change or variation, and thus make a comparison between them. Natural variation is measured by the "intravarianza" or "error variance". If samples are from the same population, the variance of the means is the umpteenth part of the population variance. If that situation is ruled out, it is because, apart from random, something else makes samples different. If these differences are only due to random, the two variances corrected by its degrees of freedom are of the same order and its quotient is worth roughly 1. The ratio of corrected variances has distribution F, so it searches in the F-distribution tables if the obtained relations are acceptable as close to 1 or not (Salvarrey, 2000).

The analysis of variance can be performed with a dependent variable (univariate ANOVA) or with multiple dependent variables (multivariate ANOVA or MANOVA). It is also possible to perform the analysis of variance (ANOVA univariate) or with several factors (multifactorial ANOVA).

The so-called univariate general linear model (MGL) performs a regression analysis and one of variance for a dependent variable by means of one or more factors or independent variables (factors divide the population into groups). With this procedure the hypotheses about the effects of one or more variables (factors or treatments) on the average of a single dependent variable can be contrasted. This analysis requires that the dependent variable be quantitative and that the categorical factors may have numeric values or string (letters).

FUNCTIONS

The function lillie.testof the package nortest (Gross, 2013) is used to make Kolmogorov-Smirnov test for normality with the correction of Lilliefors. The function shapiro.test is used to perform the Shapiro-Wilk test, the linear model is done with the function lm and the chart of interaction with the function interaction.plot, the three of the base stats package. The ANOVA is done with the function Anova of the package "car" (Fox et al., 2014). For the posthoc tests the function glht of the multcomp package is used (Hothorn et al. , 2014). For the beanplot, the function beanplot of the package beanplot is used (Kampstra, 2008;) Kampstra, 2013). The Levene test is done with the function levene.test of the package lawstat (Gastwirth et al., 2013). The asymmetry and kurtosis are performed with the functions skewness and kurtosis, respectively, of package e1071 (Meyer et al., 2014).

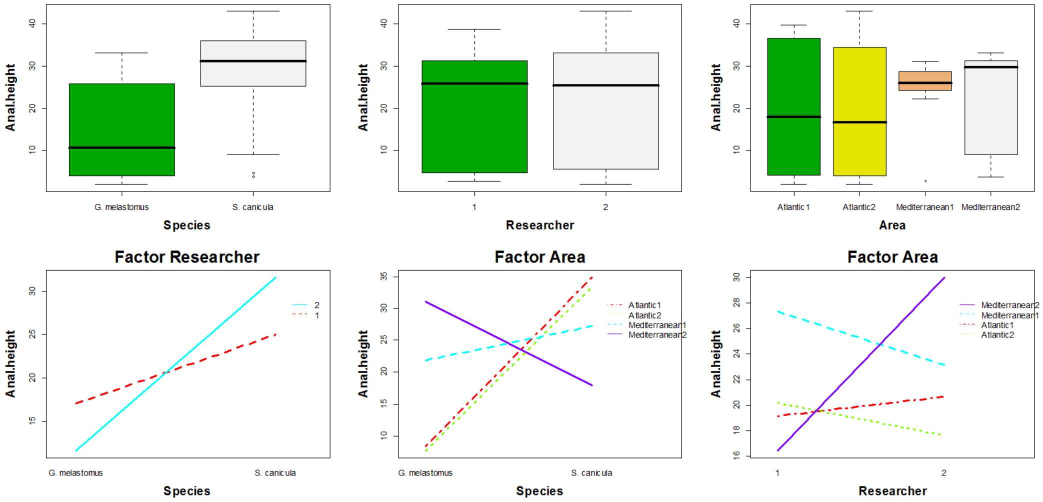
EXAMPLE

Biometric data of two species of sharks (*Scyliorhinus canicula* and *Galeus melastomus*), which were taken by two researchers in four areas (two in the Mediterranean and two in the Atlantic). The objective is to determine if there are significant differences in the length of the anal fin between species, areas and researchers.

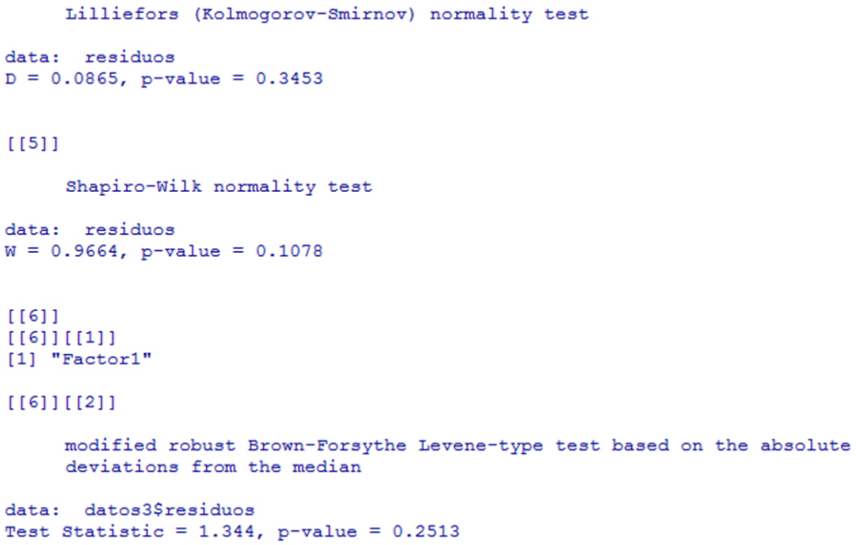
In the script the argument *mfrow=c(2,3)* to make a panel in which 6 graphics are put in 2 rows and 3 columns is specified.

Figure IX.4 shows the panel with the medium and distribution of the values of the length of the anal fin among species, researchers and areas, as well as the graphics of interaction between all factors.

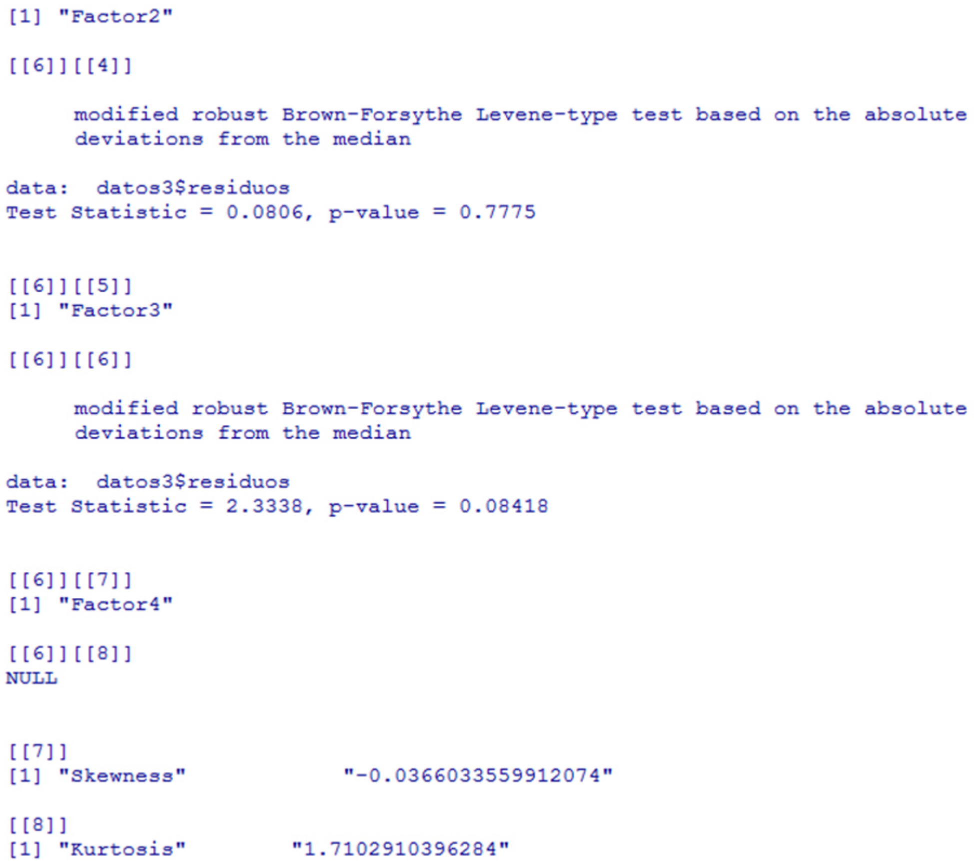
Figure IX.4. Length of the anal fin among species, researchers and areas, in addition to the graphics of interaction between factors.



The residuals fit a Normal distribution: Kolmogorov-Smirnov test with the Lilliefors correction (p = 0.345) and Shapiro-Wilk (p = 0.108), and there is homogeneity of variances for the factor 1 (p = 0.251), for the factor 2 (p = 0.777) and for the factor 3 (p = 0.084). Therefore, it is not necessary to transform the data.

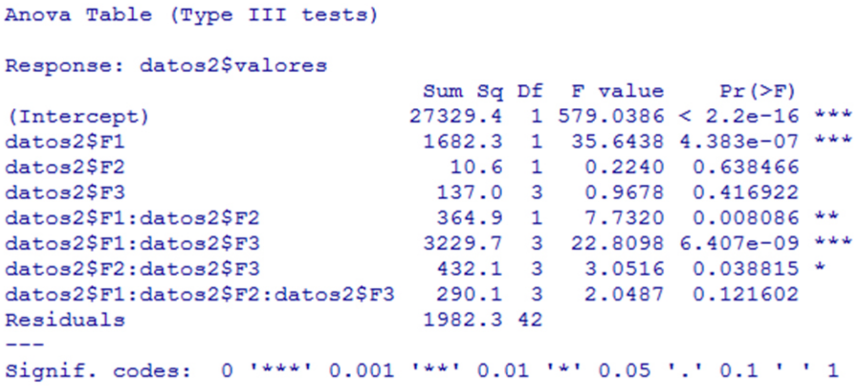


```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0865, p-value = 0.3453


[[5]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9664, p-value = 0.1078


[[6]]
[[6]][[1]]
[1] "Factor1"

[[6]][[2]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$residuos
Test Statistic = 1.344, p-value = 0.2513
```



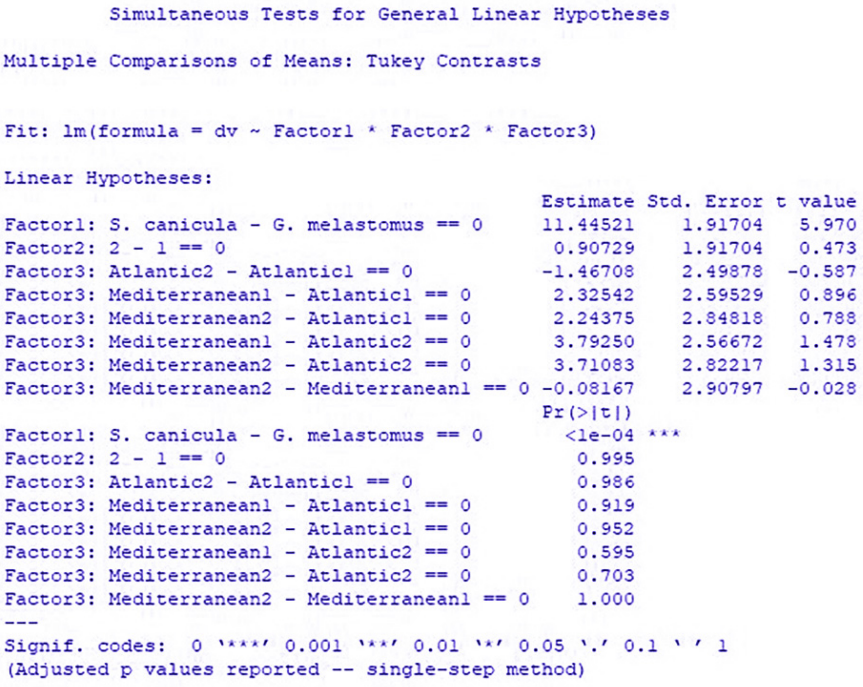
```
[1] "Factor2"

[[6]][[4]]

      modified robust Brown-Forsythe Levene-type test based on the absolute
      deviations from the median

data:  datos3$residuos
Test Statistic = 0.0806, p-value = 0.7775


[[6]][[5]]
[1] "Factor3"

[[6]][[6]]

      modified robust Brown-Forsythe Levene-type test based on the absolute
      deviations from the median

data:  datos3$residuos
Test Statistic = 2.3338, p-value = 0.08418


[[6]][[7]]
[1] "Factor4"

[[6]][[8]]
NULL


[[7]]
[1] "Skewness"             "-0.0366033559912074"

[[8]]
[1] "Kurtosis"         "1.7102910396284"
```

Despite considering the effect «Researcher» and «Area», there are clear differences in the length of the anal fin among species (ANOVA, $F_{1,42}$ = 35.6, p < 0.001). However, there are no significant differences between researchers (ANOVA, $F_{1,42}$= 0.22, p = 0.638), i.e., there are no differences in criteria when taking the measurement. There are differences between zones (ANOVA, $F_{3,42}$= 0.97, p = 0.417). The ANOVA table also shows the combined effect of the factors, considered as an additional factor which includes the possibility that - in addition to the effect of the species, researcher and the area, which added - appears another, which also adds to the previous ones, which is the interaction between them. It is noted that some interactions are significant, as «Species*Researcher» (ANOVA, $F_{1,42}$= 7.73, p = 0.008). This is observed in the graph of interaction between the factors species and researcher (Figure IX.4, figure on the left of the second row), since the lines are not parallel, but intersect, which is indicating that the researcher 1 gets a larger size of anal fin for the species *G. melastomus* than the researcher 2, while on the other hand, the researcher 2 gets a larger size of anal fin for the species *S. canicula* than the researcher 1, that is to say, the researchers measured differently the species. However, an interaction should not be taken into account when one of the factors that comprise it are not significant, in this case, a researcher factor.



```
Anova Table (Type III tests)

Response: datos2$valores
                               Sum Sq Df  F value    Pr(>F)    
(Intercept)                   27329.4  1 579.0386 < 2.2e-16 ***
datos2$F1                      1682.3  1  35.6438 4.383e-07 ***
datos2$F2                        10.6  1   0.2240  0.638466    
datos2$F3                       137.0  3   0.9678  0.416922    
datos2$F1:datos2$F2             364.9  1   7.7320  0.008086 ** 
datos2$F1:datos2$F3            3229.7  3  22.8098 6.407e-09 ***
datos2$F2:datos2$F3             432.1  3   3.0516  0.038815 *  
datos2$F1:datos2$F2:datos2$F3   290.1  3   2.0487  0.121602    
Residuals                      1982.3 42                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As in the case of the factor «Area» there are 4 levels, this could lead to perform tests *post hoc* to find out what are the different levels. However, it makes no sense to make comparisons with a factor that is not significant, because if a single comparison were significant, it would also be the factor. In any case R shows the *post hoc* results of the Tukey test, even for those factors that have only 2 levels. By observing the Factor 3, which is the «Area», as expected, it is noted that there are no significant differences when comparing each of the areas: area 2 with 1 (p = 0.986), area 3 with 1 (p = 0.919), area 4 with 1 (p = 0.952), area 3 with 2 (p = 0.595), area 4 with 2 (p = 0.703), and area 4 with 3 (p = 1).



```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = dv ~ Factor1 * Factor2 * Factor3)

Linear Hypotheses:
                                              Estimate Std. Error t value
Factor1: S. canicula - G. melastomus == 0     11.44521    1.91704   5.970
Factor2: 2 - 1 == 0                            0.90729    1.91704   0.473
Factor3: Atlantic2 - Atlantic1 == 0           -1.46708    2.49878  -0.587
Factor3: Mediterranean1 - Atlantic1 == 0       2.32542    2.59529   0.896
Factor3: Mediterranean2 - Atlantic1 == 0       2.24375    2.84818   0.788
Factor3: Mediterranean1 - Atlantic2 == 0       3.79250    2.56672   1.478
Factor3: Mediterranean2 - Atlantic2 == 0       3.71083    2.82217   1.315
Factor3: Mediterranean2 - Mediterranean1 == 0 -0.08167    2.90797  -0.028
                                              Pr(>|t|)    
Factor1: S. canicula - G. melastomus == 0       <1e-04 ***
Factor2: 2 - 1 == 0                              0.995    
Factor3: Atlantic2 - Atlantic1 == 0              0.986    
Factor3: Mediterranean1 - Atlantic1 == 0         0.919    
Factor3: Mediterranean2 - Atlantic1 == 0         0.952    
Factor3: Mediterranean1 - Atlantic2 == 0         0.595    
Factor3: Mediterranean2 - Atlantic2 == 0         0.703    
Factor3: Mediterranean2 - Mediterranean1 == 0    1.000    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Value

A TXT file with the results of the ANOVA is obtained, a boxplot or a beanplot for each factor can be displayed, in addition to displaying a graph of interaction between all the factors.

References

Azzimonti, J.C. (2003) *Bioestadística aplicada a Bioquímica y Farmacia*. Universidad Nacional de Misiones. Editorial Universitaria, Argentina.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Gastwirth, J.L., Gel, Y.R., Hui, W.L.W., Lyubchich, V., Miao, W. & Noguchi, K. (2013). An R package for biostatistics, public policy, and law. R package version 2.4.1. Available at: http://CRAN.R-project.org/package=lawstat.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Hothorn, T., Bretz, F., Westfall, P, Heiberger, R.M. & Schuetzenmeister, A. (2014) Simultaneous Inference in General Parametric Models. R package version 1.3-3. Available at: http://CRAN.R-project.org/package=multcomp.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. & Lin, C.C (2014) Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. Available at: http://CRAN.R-project.org/package=e1071.

Salvarrey, L. (2000) *Curso de estadística básica*. Universidad de la República del Uruguay.

Examples

```
## Not run:

data(ZIX3)

#Comparison of the length of the anal fin between two shark species,
#measured by two researchers and in several sampling areas

IX3(data=ZIX3, variables="Anal.height", Factor1="Species", Factor2="Researcher",
Factor3="Area", mfrow=c(2,3))


## End(Not run)
```

IX4 *ANALYSIS OF COVARIANCE (ANCOVA)*

Description

An analysis of covariance with optional graphics is applied.

Usage

```
IX4(data, variables,  Factor1, Factor2=NULL, Factor3=NULL, Covariable1,
Covariable2=NULL, Covariable3=NULL, Covariable4=NULL, SS=3, trans=NULL,
ResetPAR=TRUE, mfrow=NULL,  CEXPCH=1.3, graph1=TRUE, XLAB=NULL, YLAB=NULL,
PAR1=NULL, COLOR1=NULL, PCH1=NULL,  MAIN1=NULL, LEGEND1=NULL, AXIS1=NULL,
MTEXT1= NULL, TEXT1=NULL, graph2=TRUE, PAR2=NULL, COLOR2=NULL, PCH2 =NULL,
MAIN2=NULL, LEGEND2=NULL, AXIS2=NULL, MTEXT2= NULL, TEXT2=NULL, graph3=TRUE,
PAR3=NULL, COLOR3=NULL,  PCH3=NULL, MAIN3=NULL, LEGEND3=NULL, AXIS3=NULL,
MTEXT3= NULL, TEXT3=NULL, INTPLOT1.2=TRUE, INTPLOT1.3=TRUE, INTPLOT2.3=TRUE,
INTERPLOT1.2=NULL, INTERPLOT1.3=NULL, INTERPLOT2.3=NULL, OrderCat1=NULL,
LabelCat1=NULL, OrderCat2=NULL, LabelCat2=NULL, file="Output.txt")
```

Arguments

`data`	Data file.
`variables`	Dependent variable.
`Factor1`	First factor.
`Factor2`	Second factor.
`Factor3`	Third factor.
`Covariable1`	First covariate.
`Covariable2`	Second covariate.
`Covariable3`	Third covariate.
`Covariable4`	Fourth covariate.
`SS`	When there are several factors, the decomposition of the sum of squares in parts attributed to each one of the factors is not unique. Four types of sums of squares are commonly used: type I for nested models, type II for balanced models, type III for unbalanced models (also balanced), and type IV for models with empty boxes. The type III is suitable for most applications, so it will be the one used by default. This description is very simplified; interested readers should consult supplementary bibliography for a correct use of the ANOVA with complex models.
`trans`	Type of transformation that is applied to the data: 1. NULL (untransformed) 2. 1/x2 3. 1/x

4. LN
5. LOG
6. SQR (square root)
7. x2
8. x3
9. EXP (exponential)
10. ASN (arcsine)

`ResetPAR` If FALSE the conditions are not placed by default in the PAR function and those defined by the user in previous graphics are kept.

`mfrow` If it is NULL and there are various graphics, these appear in separate windows. If the graphics go into panels, this argument is a vector with the format c(nr, nc) indicating the number of figures per row (nr) and column (nc), by first filling the rows.

`CEXPCH` Size of the graphic symbols.

`graph1` If NULL, the graph of factor 1 is not performed with the covariate 1.

`XLAB` Legend of the X axis.

`YLAB` Legend of the Y axis.

`PAR1` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 1.

`COLOR1` This allows to modify the colors of the graph 1, but should be as many different groups as the factor 1 has.

`PCH1` Vector with the symbols of the graphic 1. If NULL, they are automatically calculated starting with the symbol 15.

`MAIN1` Title of the graph of factor 1.

`LEGEND1` It allows to modify the legend on the graph of factor 1.

`AXIS1` It allows to add axes to the graph of factor 1.

`MTEXT1` It allows to add texts on the margins of the graph of factor 1.

`TEXT1` It allows to add text in any area of the inner part of the graph of factor 1.

`graph2` If NULL, the graph of factor 2 is not performed with the covariate 1.

`PAR2` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 2.

`COLOR2` This allows to modify the colors of the graph 2, but should be as many different groups as the factor 2 has.

`PCH2` Vector with the symbols of graphic 2. If NULL, they are automatically calculated starting with the symbol 15.

`MAIN2` Title of the graph of the factor 2.

`LEGEND2` It allows to modify the legend on the graph of the factor 2.

`AXIS2` It allows to add axes to the graph of the factor 2.

`MTEXT2` It allows to add text in the margins of the graph of the factor 2.

`TEXT2` It allows to add text in any area of the inner part of the graph of the factor 2.

`graph3` If NULL, the graph of factor 3 is not performed with the covariate 1.

`PAR3` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 3.

`COLOR3` This allows to modify the colors of the graph 3, but should be as many different groups as the factor 3 has.

`PCH3` Vector with the symbols of graphic 3. If NULL, they are automatically calculated starting with the symbol 15.

`MAIN3` Title of the graph of the factor 3.

`LEGEND3` It allows to modify the legend on the graph of factor 3.

`AXIS3` It allows to add axes to the graph of the factor 3.

`MTEXT3` It allows to add text in the margins of the graph of the factor 3.

`TEXT3` It allows to add text in any area of the inner part of the graph of the factor 3.

`INTPLOT1.2` If TRUE, it shows the graph of interaction of factor 1 with the factor 2.

`INTPLOT1.3` If TRUE, it shows the graph of interaction of factor 1 with the factor 3.

`INTPLOT2.3` If TRUE, it shows the graph of interaction of factor 2 with the factor 3.

`INTERPLOT1.2` It allows to specify the characteristics of the graph of the interaction of factor 1 with factor 2.

`INTERPLOT1.3` It allows to specify the characteristics of the graph of the interaction of factor 1 with factor 3.

`INTERPLOT2.3` It allows to specify the characteristics of the graph of the interaction of factor 2 with factor 3.

`OrderCat1` It allows to specify a vector with the order in which the categories of factor 1 are shown in the graph of interaction.

`LabelCat1` It allows to specify a vector with the names of the categories of factor 1 in the graph of interaction.

`OrderCat2` It allows to specify a vector with the order in which the categories of factor 2 are shown in the graph of interaction.

`LabelCat2` It allows to specify a vector with the names of the categories of factor 2 in the graph of interaction.

`file` TXT FILE. Name of the output file with the results.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.1.5. Analysis of covariance (ANCOVA)

The analysis of covariance (ANCOVA) is a statistical technique that aims to eliminate the effect of possible variables that can be related to the dependent variable.

To control these variables, a variance analysis is performed, in which the dependent variable is the error in the forecasts, or residual, performing a linear regression analysis with covariates as independent variables and the dependent variable is equal to that of our initial ANOVA.

In ANCOVA, the interpretation of the results, as in the ANOVA, was based on the effects of the factors on our dependent variables and the interactions among the factors studied. According to Steel & Torrie (1985) this analysis can be used for:

1. Handling errors and increase the accuracy of our analysis.

2. Adjusting the means of the dependent variables to the differences with the values of the independent variables used.

3. Assist in the interpretation of our data, especially in regard to the nature of the effects of the treatments.

4. Partitioning of a total covariance or sum of cross products in components.

5. Estimating missing data.

In order to interpret an ANCOVA correctly, it is necessary to do the ANOVA first, i.e., do the analysis only taking into account the dependent variable and the factor(s), without considering the covariate(s). This is due to the fact that the interpretation of the ANCOVA differs depending on whether the results obtained in the ANCOVA are the same or different from those obtained in the ANOVA.

In the ANCOVA any variable that does not have a significant effect can be eliminated from the analysis. If all the covariates used have no significant effects, the conclusions of the ANCOVA and ANOVA should be very similar.

If one or more covariates have significant effects, two situations can occur:

1. The outcome of ANOVA and ANCOVA is the same. This means that the effect of factors on the dependent variable is not modified and, therefore, the possible effect of the(s) covariate(s) on the dependent variable does not affect the relationship of the latter with the factors.

2. If the result of the ANOVA and ANCOVA is different, it can be due to two reasons:

2a. It may be because a significant ANOVA effect becomes insignificant and, thus, the effect detected in the ANOVA is due to the effect of the(s) covariate(s) and not to the(s) independent variable(s)(factors).

2b. It can also be because an insignificant effect ANOVA becomes significant, which indicates that the factor, even not being related to the dependent variable globally considered, correlates with the part of the dependent variable not explained or not attributable to the covariate(s).

FUNCTIONS

The function lillie.test of the nortest package (Gross, 2013) is used to perform the Normality test of Kolmogorov-Smirnov with Lilliefors' correction.

The function shapiro.test is used to perform the Shapiro-Wilk test, the linear model is performed with the function lm and the chart of interaction with the function interaction.plot, all three of the base stats package.

The ANCOVA is performed with the function Anova of the car package (Fox et al., 2014). The Levene test is performed with the function levene.test package of the lawstat package (Gastwirth et al., 2013). The asymmetry and kurtosis are performed with the functions skewness and kurtosis, respectively, of the package e1071 (Meyer et al., 2014).

EXAMPLE

This compares the length of the anal fin between two species of sharks, as measured by two researchers and in several areas of sampling considering the total length of the body as a covariate.

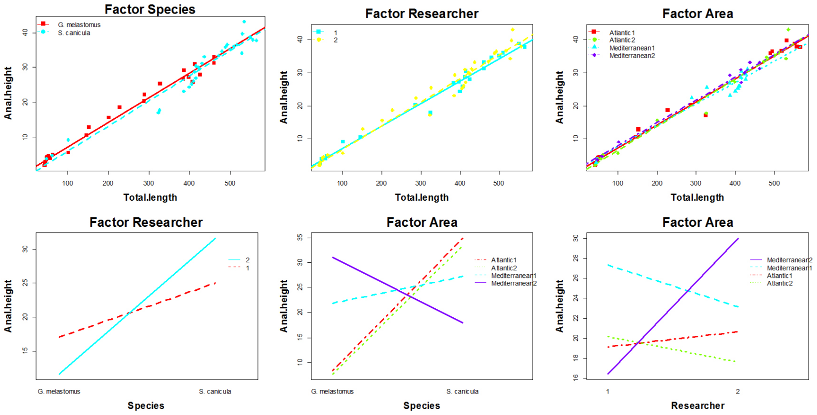
Step 1.

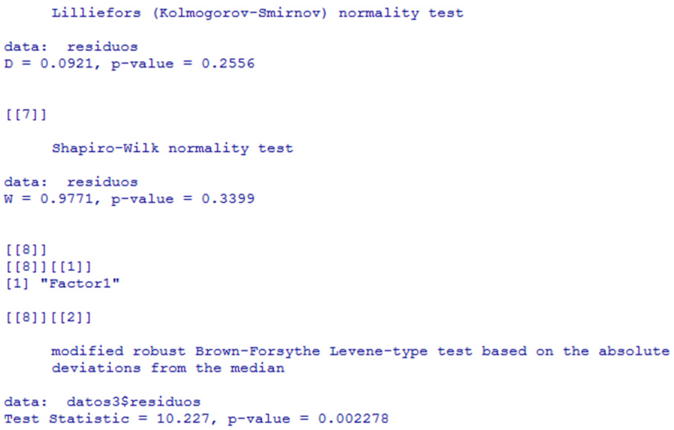
Figure IX.5 shows the panel with the relationship between total length and the length of the anal fin, differentiating between species, researchers and the areas, as well as the graphics of interaction between factors.

First we note the normality of residuals and the homogeneity of variances. It is noted that both the Kolmogorov-Smirnov test with the correction of Lilliefors (p = 0.256) as the Shapiro-Wilk (p = 0.34) show values of p > 0.05 and, therefore, it is accepted the null hypothesis of goodness of fit and the residuals comply with a Normal distribution.

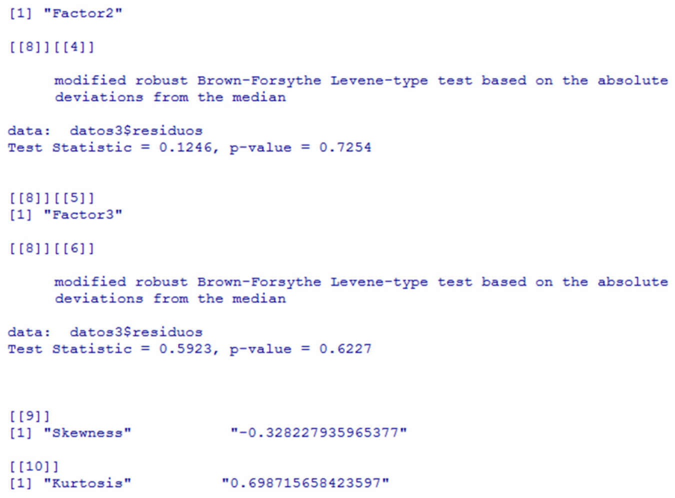
The Levene test, considering the median, shows that the homogeneity of variances for the factor 1 (p = 0.002), is not fulfilled, but is true for 2 factor (p = 0.725) and factor 3 (p = 0.623). Therefore, it is necessary to transform the data.

Figure IX.5. Relationship between the total length and the length of the fin between species, researchers and areas, along with the graphics of interaction between factors.





```
      Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0921, p-value = 0.2556


[[7]]

      Shapiro-Wilk normality test

data:  residuos
W = 0.9771, p-value = 0.3399


[[8]]
[[8]][[1]]
[1] "Factor1"

[[8]][[2]]

      modified robust Brown-Forsythe Levene-type test based on the absolute
      deviations from the median

data:  datos3$residuos
Test Statistic = 10.227, p-value = 0.002278
```

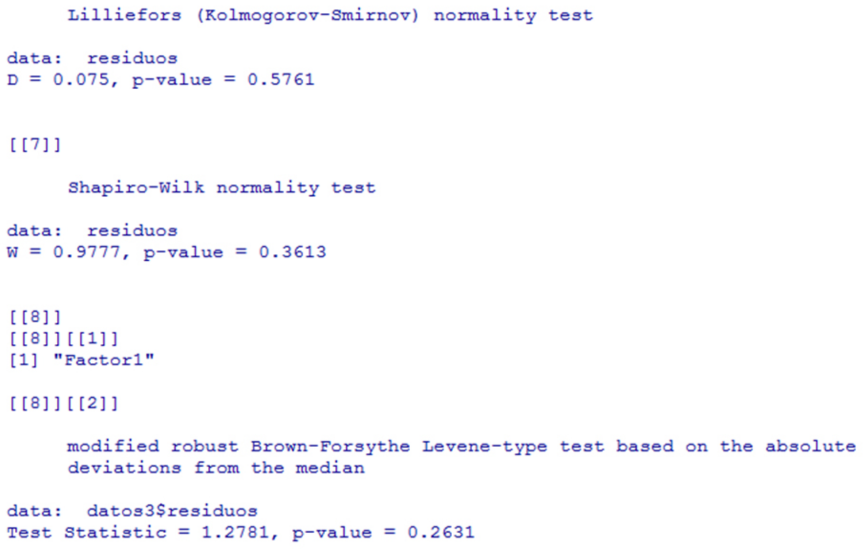


```
[1] "Factor2"

[[8]][[4]]

      modified robust Brown-Forsythe Levene-type test based on the absolute
      deviations from the median

data:  datos3$residuos
Test Statistic = 0.1246, p-value = 0.7254


[[8]][[5]]
[1] "Factor3"

[[8]][[6]]

      modified robust Brown-Forsythe Levene-type test based on the absolute
      deviations from the median

data:  datos3$residuos
Test Statistic = 0.5923, p-value = 0.6227



[[9]]
[1] "Skewness"            "-0.328227935965377"

[[10]]
[1] "Kurtosis"           "0.698715658423597"
```

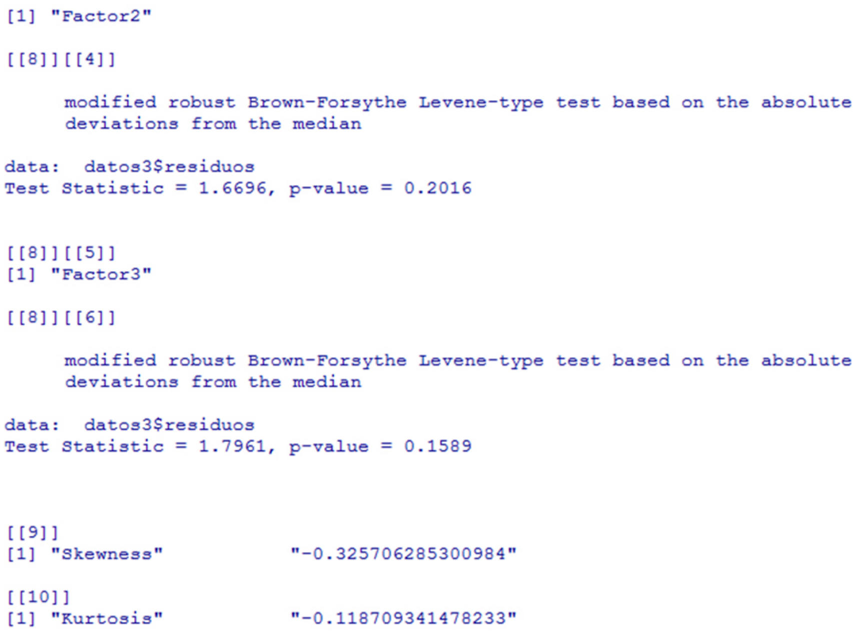
Step 2.

The analysis is repeated but transforming data with the *trans="SQR"* argument. Now it is observed that both the Kolmogorov-Smirnov test with correction of Lilliefors (p = 0.576) as the Shapiro-Wilk test (p = 0.361) have values of p > 0.05 and, therefore, the null hypothesis of goodness of fit is accepted and and the residuals comply with a Normal distribution.

The Levene test, considering the median, shows that the homogeneity of variances for the factor 1 (p = 0.263), for the factor 2 (p = 0.202) and for the factor 3 (p = 0.159) is fulfilled.

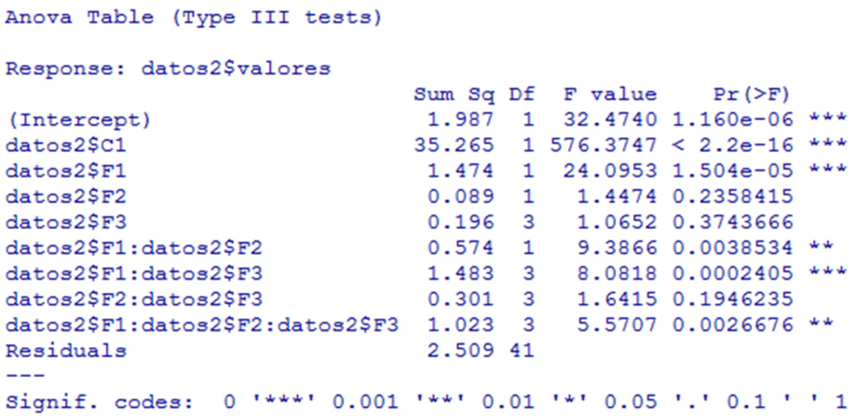


```
      Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.075, p-value = 0.5761


[[7]]

      Shapiro-Wilk normality test

data:  residuos
W = 0.9777, p-value = 0.3613


[[8]]
[[8]][[1]]
[1] "Factor1"

[[8]][[2]]

      modified robust Brown-Forsythe Levene-type test based on the absolute
      deviations from the median

data:  datos3$residuos
Test Statistic = 1.2781, p-value = 0.2631
```



```
[1] "Factor2"

[[8]][[4]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 1.6696, p-value = 0.2016


[[8]][[5]]
[1] "Factor3"

[[8]][[6]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 1.7961, p-value = 0.1589



[[9]]
[1] "Skewness"           "-0.325706285300984"

[[10]]
[1] "Kurtosis"           "-0.118709341478233"
```

The ANOVA table shows that there is a clear relationship between the total length and the length of the anal fin(ANCOVA, $F_{1,41}$ = 576.37, p < 0.001). By considering the effect of total length remain clear differences in the length of the anal fin between species (ANOVA, $F_{1,41}$ = 24.09, p < 0.001), but not among researchers (ANOVA, $F_{1,41}$= 1.44, p = 0.235) or between zones (ANOVA, $F_{3,41}$= 1.06, p = 0.374).



```
Anova Table (Type III tests)

Response: datos2$valores
                           Sum Sq Df  F value    Pr(>F)    
(Intercept)                 1.987  1  32.4740 1.160e-06 ***
datos2$C1                  35.265  1 576.3747 < 2.2e-16 ***
datos2$F1                   1.474  1  24.0953 1.504e-05 ***
datos2$F2                   0.089  1   1.4474 0.2358415    
datos2$F3                   0.196  3   1.0652 0.3743666    
datos2$F1:datos2$F2         0.574  1   9.3866 0.0038534 ** 
datos2$F1:datos2$F3         1.483  3   8.0818 0.0002405 ***
datos2$F2:datos2$F3         0.301  3   1.6415 0.1946235    
datos2$F1:datos2$F2:datos2$F3 1.023  3   5.5707 0.0026676 ** 
Residuals                   2.509 41                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Value

A TXT file with the results of the ANOVA is obtained, a graph for each factor with the covariate 1 can be displayed, as well as a graph of interaction of factor 1 with each of the other factors can also be shown.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Gastwirth, J.L., Gel, Y.R., Hui, W.L.W., Lyubchich, V., Miao, W. & Noguchi, K. (2013). An R package for biostatistics, public policy, and law. R package version 2.4.1. Available at: http://CRAN.R-project.org/package=lawstat.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. & Lin, C.C (2014) Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. Available at: http://CRAN.R-project.org/package=e1071.

Steel, R.G.D. & Torrie, J.H. (1985) *Bioestadística: Principios y procedimientos*. McGraw-Hill, Bogotá.

Examples

```
## Not run:

data(ZIX3)

#Step 1
#A single graph for each factor and interaction graph between factors

IX4(data=ZIX3,  variables="Anal.height", Factor1="Species",
Factor2="Researcher", Factor3="Area", Covariable1="Total.length")

#The 6 graphics are combined into one with mfrow, distances of
#margins are specified with PAR (mar =) and the color of graph 2 is changed

IX4(data=ZIX3,  variables="Anal.height", Factor1="Species",
Factor2="Researcher",
Factor3="Area", Covariable1="Total.length", mfrow=c(2,3), PAR1=c("mar=c(5,5,2,1)",
"font.lab=2", "cex.lab=1.5"), COLOR2= c("#00FFFFFF", "#FFFF00FF"))

#Step 2

IX4(data = ZIX3 , variables="Anal.height", Factor1="Species",
Factor2="Researcher",
Factor3="Area", Covariable1="Total.length", trans = "SQR", mfrow = c(2,3),
PAR1 = c("mar=c(5,5,2,1)", "font.lab=2", "cex.lab=1.5"), COLOR2=c("#00FFFFFF","#FFFF00FF"))


## End(Not run)
```

IX5 *ANOVA FOR FACTORS WITH REPEATED MEASURES*

Description

ANOVA was applied for dependent samples and the data can be displayed with a boxplot or beanplot.

Usage

```
IX5(data, varTime, graph="Boxplot", SS=2, PAR=NULL, ResetPAR=TRUE,
YLAB="Dependent variable", XLAB="Time", LabelCat=NULL, COLOR=NULL,
BOXPLOT=NULL, BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT=NULL,
TEXT=NULL, file="Output.txt")
```

Arguments

`data` Data file.

`varTime` Columns in which the dependent variable has been measured over time.

`graph` If it is NULL, there is no graph and the other options are "Boxplot" or "Beanplot".

`SS` When there are several factors, the decomposition of the sum of squares in parts attributed to each one of the factors is not unique. Four types of sums of squares are commonly used: type I for nested models, type II for balanced models, type III for unbalanced models (also balanced), and type IV for models with empty boxes. As a model of repeated measures all the boxes of the model have equal number of data (balanced model), the type II will be used by default. This description is very simplified; interested readers should consult supplementary bibliography for a correct use of the ANOVA with complex models.

`PAR` Accessing the function PAR which allows to modify many different aspects of the chart.

`ResetPAR` If FALSE, the conditions are not placed by default in the function PAR and those defined by the user on previous graphics are maintained.

`YLAB` A legend for the Y axis in the boxplot and beanplot.

`XLAB` A legend for the X axis in the boxplot and beanplot.

`LabelCat` It allows to specify a vector with the order in which categories are displayed in the graph.

`COLOR` Vector with the color of the categories of the chart.

`BOXPLOT` It allows to specify the characteristics of the boxplot.

`BEANPLOT` It allows to specify the characteristics of the beanplot.

`LEGEND` It allows to include a legend.

`AXIS` It allows to add axes.

`MTEXT` It allows to add texts in the margins of the chart.

`TEXT` It allows to add a text in any area of the inner part of the chart.

`file` TXT FILE. Name of the output file.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.1.6. ANOVA for factors with repeated measures

The function IX2 showed that there was a form of *t*-test that was used in the case of wanting to make a contrast of homogeneity of related or matched data. However, it may be the case that the factor has more than two levels or that has more than one factor. For example, several people were selected and asked to give a score from 0 to 10 on the quality of five different perfumes. In these cases the *t*-test only allows to compare two perfumes together. To compare the five joint, an analysis of variance for repeated measures must be used.

The analysis with repeated measures is designed to more precisely determine the residual error, which will be less if we are able to control or delete the variability between subjects and leave only the variability intra subject. The smaller is the residual error, the more precise determination of significant effects, as each effect is compared with the residual variability to know if it is significant.

This type of analysis of variance can be univariate or multifactorial, i.e., the fact of working with one or more factors with repeated measures in all factors or only in one of them. This also allows to include covariates, i.e., to run an ANCOVA (analysis of covariance, see function IX7).

FUNCTIONS

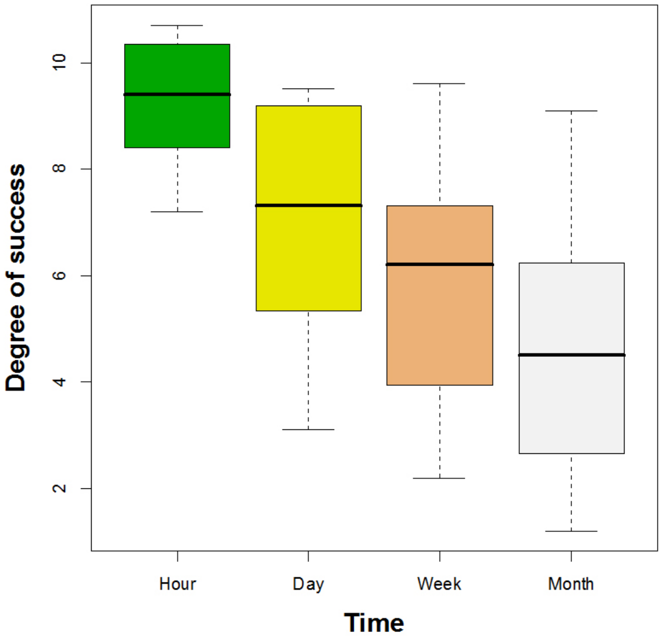
For the beanplot, the function beanplot of the package beanplot (Kampstra, 2008; Kampstra, 2013) is used. The linear model is performed with the functionlm of the package stats. The ANOVA is performed with the Anova of the package car (Fox et al., 2014).

EXAMPLE

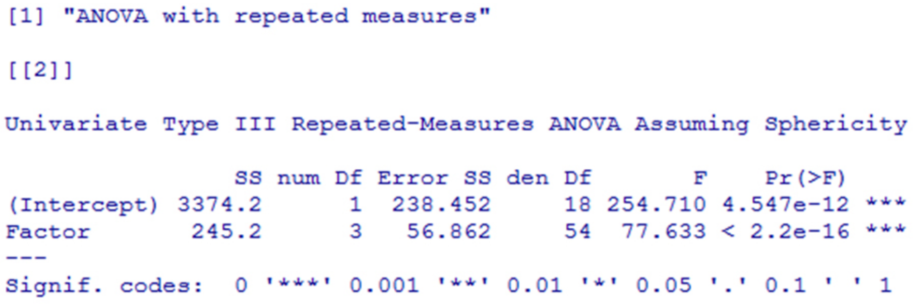
In an experiment conducted with expert tasters and people who had no experience of tasting, they were taught to identify 15 types of wines from different regions. The objective was to determine whether the ability to ascertain the provenance of the wine varied over time (after one hour, one day, one week and one month). To do this, each time each person evaluated a large number of samples and recorded the degree of success on a scale of 0 to 12.

Figure IX.6 clearly shows how the skill of the tasters decreases with time.

Figure IX.6. Degree of success of the tasters over time.

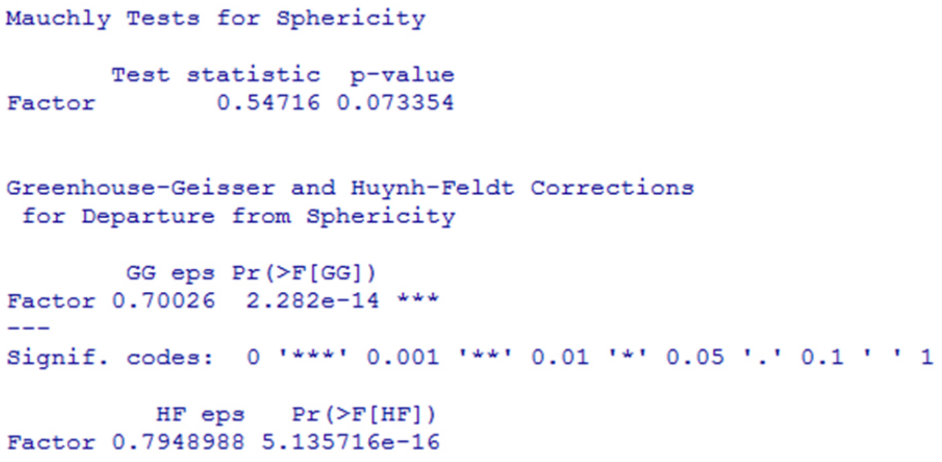


The results indicate that the time is a significant factor (p < 0.001). The observed differences in the degree of success over time are not due to chance, but that the learned capacity to correctly identify the wines is lost with time.



```
[1] "ANOVA with repeated measures"

[[2]]

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

                SS num Df Error SS den Df       F    Pr(>F)
(Intercept) 3374.2      1  238.452     18 254.710 4.547e-12 ***
Factor       245.2      3   56.862     54  77.633 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

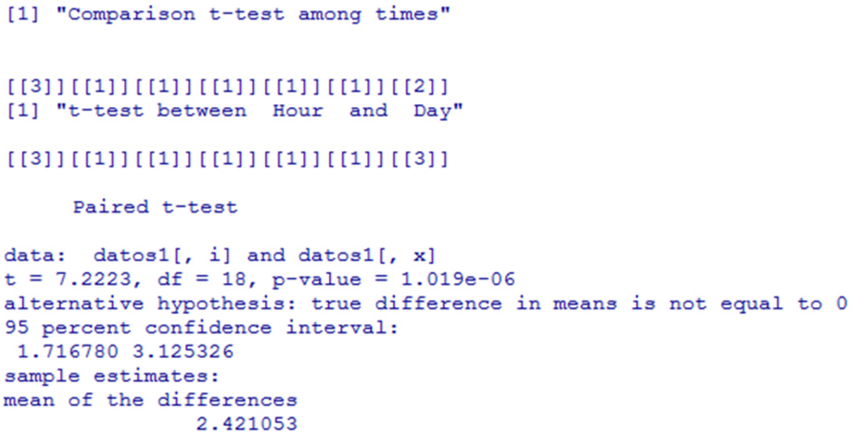
A requirement of the statistical *F* in the ANOVA with repeated measures is that the variances of the differences between each two levels of the intra-subject factor, time, must be equal, and the levels independent of each other. This assumption is equivalent to stating that the variance-covariance matrix is proportional to the unit matrix (equal variances and null covariances), or that the joint dispersion is spherical. The table shows the results of the test of sphericity of Mauchley. As the value *p* is greater than 0.05 (p = 0.073) the assumption of sphericity is accepted.



```
Mauchly Tests for Sphericity

       Test statistic  p-value
Factor        0.54716 0.073354


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

        GG eps Pr(>F[GG])
Factor 0.70026  2.282e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          HF eps   Pr(>F[HF])
Factor 0.7948988 5.135716e-16
```

If the requirement of sphericity might not be fulfilled (if the value p is less than 0.05), one of the two modifications that are also shown should be used: Greenhouse-Geisser and Huynh-Feldt, to correct the degrees of freedom downward which allows to adjust (increase) the p value of the ANOVA contrast. In this case both corrected p values $2.282 * 10^{-14}$ and $5.136 * 10^{-16}$ are virtually zero, so it also shows a significant relationship between time and accuracy.

Once tested the significant relationship, it is necessary to know which of the levels are different (at least two are). For this a t test with the two levels to be compared must be used. The function automatically performs all the comparisons between each time. For example, the results of the comparison of the degree of coincidence after one hour and after one day are displayed. As the p-value < 0.001, one can conclude that the degree of success is significantly different between hour and day.



```
[1] "Comparison t-test among times"


[[3]][[1]][[1]][[1]][[1]][[1]][[2]]
[1] "t-test between  Hour  and  Day"

[[3]][[1]][[1]][[1]][[1]][[1]][[3]]

     Paired t-test

data:  datos1[, i] and datos1[, x]
t = 7.2223, df = 18, p-value = 1.019e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.716780 3.125326
sample estimates:
mean of the differences
               2.421053
```

When making multiple comparisons, such as the above-mentioned, it must be remembered that it is necessary to use some type of protection, especially when making many comparisons, to avoid the increase of the probability of error of type I. For example, one can divide the significance level by the number of comparisons (Bonferroni method).

Value

A TXT file with the results of the ANOVA is obtained for dependent samples and a boxplot or beanplot can also be displayed.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: `http://CRAN.R-project.org/package=car`.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: `http://CRAN.R-project.org/package=beanplot`.

Examples

```
## Not run:

data(ZIX5)

#Variation in the ability to ascertain the origin of a
#wine versus time

IX5(data = ZIX5 , varTime = c("Hour","Day","Week","Month") ,
YLAB = "Degree of success")


## End(Not run)
```

IX6 *ANOVA FOR FACTORS WITH REPEATED AND NON-REPEATED MEASURES*

Description

An ANOVA in which there are factors with repeated measures (factors intra-subject) and factors without repeated measures (factors inter-subject) is applied. In addition to displaying the data with a boxplot or beanplot and an interaction graph of each of the factors along with the time factor.

Usage

```
IX6(data, varTime, Factor1, Error, Factor2=NULL, Factor3=NULL, SS=2,
ResetPAR=TRUE, mfrow=NULL, YLAB="Dependent variable", XLAB="Time",
graphT="Boxplot", PART=NULL, LabelCatT=NULL, COLORT=NULL, BOXPLOTT=NULL,
BEANPLOTT=NULL, LEGENDT=NULL, AXIST=NULL, MTEXTT= NULL, TEXTT=NULL,
graph1="Boxplot", PAR1=NULL, OrderCat1=NULL, LabelCat1=NULL, COLOR1=NULL,
BOXPLOT1=NULL, BEANPLOT1=NULL, LEGEND1=NULL, AXIS1=NULL, MTEXT1= NULL,
TEXT1=NULL, graph2="Boxplot",  PAR2=NULL, OrderCat2=NULL, LabelCat2=NULL,
COLOR2=NULL, BOXPLOT2=NULL, BEANPLOT2=NULL, LEGEND2=NULL, AXIS2=NULL,
MTEXT2= NULL, TEXT2=NULL, graph3="Boxplot", PAR3=NULL, OrderCat3=NULL,
LabelCat3=NULL, COLOR3=NULL, BOXPLOT3=NULL, BEANPLOT3=NULL, LEGEND3=NULL,
AXIS3=NULL, MTEXT3= NULL, TEXT3=NULL, INTERPLOT1=NULL, INTERPLOT2=NULL,
INTERPLOT3=NULL, file="Output.txt")
```

Arguments

`data`	Data file.
`varTime`	Columns in which the dependent variable has been measured over time.
`Factor1`	First factor.
`Error`	It is the variable that encodes the subjects studied.
`Factor2`	Second factor.
`Factor3`	Third factor.
`SS`	When there are several factors, the decomposition of the sum of squares in parts attributed to each one of the factors is not unique. Four types of sums of squares are commonly used: type I for nested models, type II for balanced models, type III for unbalanced models (also balanced), and type IV for models with empty boxes. As a model of repeated measures all the boxes of the model have equal number of data (balanced model), the type II will be used by default. This description is very simplified; interested readers should consult supplementary bibliography for a correct use of the ANOVA with complex models.
`ResetPAR`	If FALSE, the conditions are not placed by default in the function PAR and those defined by the user on previous graphics are maintained.
`mfrow`	If it is NULL and there are various graphics, these appear in separate windows. If the graphics go into panels, this argument is a vector with the format c(nr, nc) indicating the number of figures per row (nr) and column (nc), by first filling the rows.
`YLAB`	A legend for the Y axis in the boxplot and beanplot.
`XLAB`	A legend for the X axis in the boxplot and beanplot.
`graphT`	If it is NULL, there is no graph of the time factor and the other options are "Boxplot" or "Beanplot".
`PART`	This accesses the function PAR which allows to modify many different aspects of the graph of the time factor.
`LabelCatT`	This allows to specify a vector with the names of the categories of the graph of the time factor.
`COLORT`	Vector with the color of the categories of the graph of the time factor.
`BOXPLOTT`	It allows to specify the characteristics of the boxplot graph of the time factor.
`BEANPLOTT`	It allows to specify the characteristics of the beanplot graph of the time factor.
`LEGENDT`	It allows to include a legend to the graph of the time factor.
`AXIST`	It allows to add axes to the graph of the time factor.

`MTEXTT` It allows to add texts in the margins of the graph of the time factor.

`TEXTT` It allows to add text in any area of the inner part of the graph of the time factor.

`graph1` If NULL, the graphic of factor 1 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR1` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 1.

`OrderCat1` It allows to specify a vector with the order in which the categories are shown in graph 1.

`LabelCat1` It allows to specify a vector with the names of the categories of graph 1.

`COLOR1` Vector with the color of the categories of graph 1.

`BOXPLOT1` It allows to specify the characteristics of the boxplot of factor 1.

`BEANPLOT1` It allows to specify the characteristics of the beanplot of factor 1.

`LEGEND1` It allows to include a legend to the graph of factor 1.

`AXIS1` It allows to add axes to the graph of factor 1.

`MTEXT1` It allows to add texts on the margins of the graph of factor 1.

`TEXT1` It allows to add text in any area of the inner part of the graph of factor 1.

`graph2` If NULL, the graphic of factor 2 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR2` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 2.

`OrderCat2` It allows to specify a vector with the order in which the categories are shown in graph 2.

`LabelCat2` It allows to specify a vector with the names of the categories of graph 2.

`COLOR2` Vector with the color of the categories of graph 2.

`BOXPLOT2` It allows to specify the characteristics of the boxplot of factor 2.

`BEANPLOT2` It allows to specify the characteristics of the beanplot of factor 2.

`LEGEND2` It allows to include a legend to the graph of the factor 2.

`AXIS2` It allows to add axes to the graph of the factor 2.

`MTEXT2` It allows to add text in the margins of the graph of the factor 2.

`TEXT2` It allows to add text in any area of the inner part of the graph of the factor 2.

`graph3` If NULL, the graphic of factor 3 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR3` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 3.

`OrderCat3` It allows to specify a vector with the order in which the categories are shown in graph 3.

`LabelCat3` It allows to specify a vector with the names of the categories of graph 3.

`COLOR3` Vector with the color of the categories of graph 3.

`BOXPLOT3` It allows to specify the characteristics of the boxplot of factor 3.

`BEANPLOT3`	It allows to specify the characteristics of the beanplot of factor 3.
`LEGEND3`	It allows to include a legend to the graph of the factor 3.
`AXIS3`	It allows to add axes to the graph of the factor 3.
`MTEXT3`	It allows to add text in the margins of the graph of the factor 3.
`TEXT3`	It allows to add text in any area of the inner part of the graph of the factor 3.
`INTERPLOT1`	It allows to specify the characteristics of the graph of interaction of factor 1.
`INTERPLOT2`	It allows to specify the characteristics of the graph of interaction of factor 2.
`INTERPLOT3`	It allows to specify the characteristics of the graph of interaction of factor 3.
`file`	TXT FILE. Output file name.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.1.7. ANOVA for factors with repeated and non- repeated measures

See section *details* of the function IX5 or Guisande et al. (2011) for an explanation on this type of ANOVA.

FUNCTIONS

For the beanplot, the function beanplot of the package beanplot (Kampstra, 2008; Kampstra, 2013) is used.

The linear model is performed with the function lm and the graph of interaction is performed with the function interaction.plot, both with the same package stats.

The ANOVA is performed with the function Anova of the package car (Fox et al., 2014).

EXAMPLE

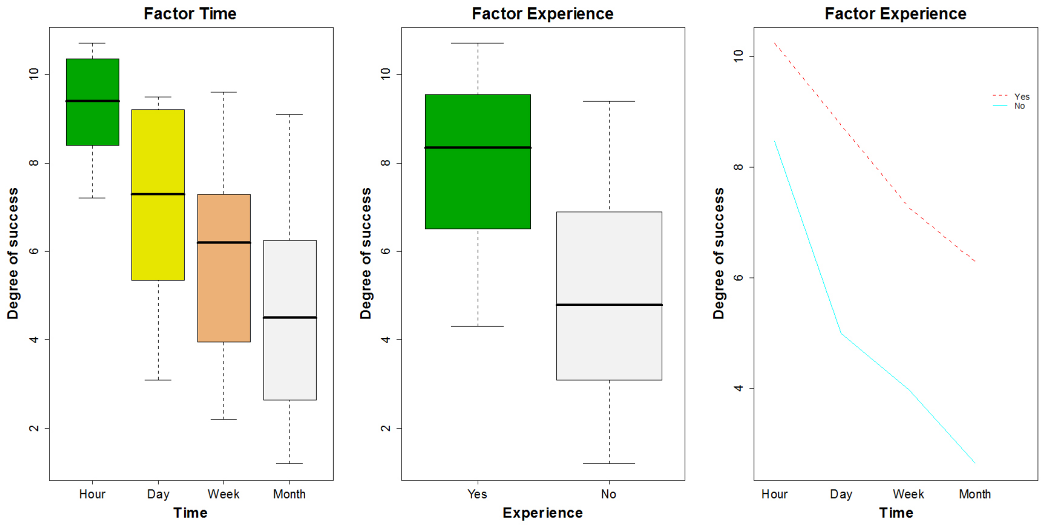
In an experiment conducted with expert tasters and people who had no experience of tasting, they were taught to identify 15 types of wines from different regions. The objective was to determine whether the ability to ascertain the provenance of the wine varied over time (after one hour, one day, one week and one month), and depending on the experience. To do this, every time each person assessed a large number of samples and the degree of success on a scale of 0 to 12 was recorded.

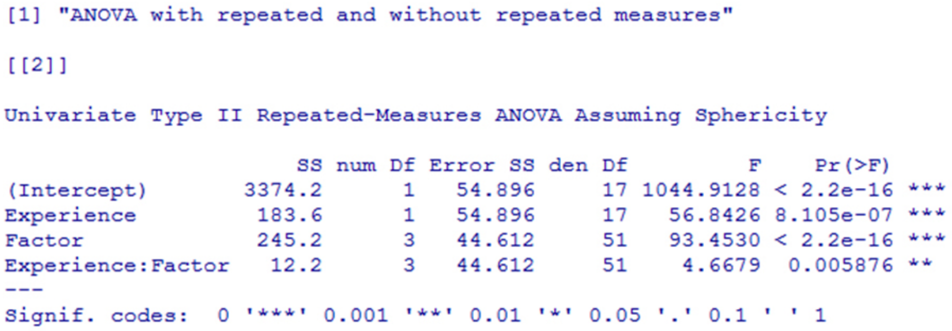
Figure IX.7 clearly shows how the degree of accuracy of the tasters decreases over time (left panel), but experienced tasters have a greater degree of success than those who do not have (central panel) and finally the interaction graph seems to show that the degree of success, or intensity of memory, decreases over time from one hour to one month, similar in both groups (both lines are approximately parallel, especially between one day and one month).

The intensity is smaller as time passes, and experts always maintain a degree of success two or three points higher than non-experts. However, it is necessary to check whether this difference is statistically significant.

The «Experience» factor is significant, with $p < 0.001$. The factor «Time» is also significant($p < 0.001$), as we had already checked, and the interaction between them, with $p = 0.006$, which means that the evolution of the memory is different in both groups, and the curves are not really parallel as it had seemed before. Therefore, the ability to identify the origin of wine decreases with different intensity over time between experts and non-experts.

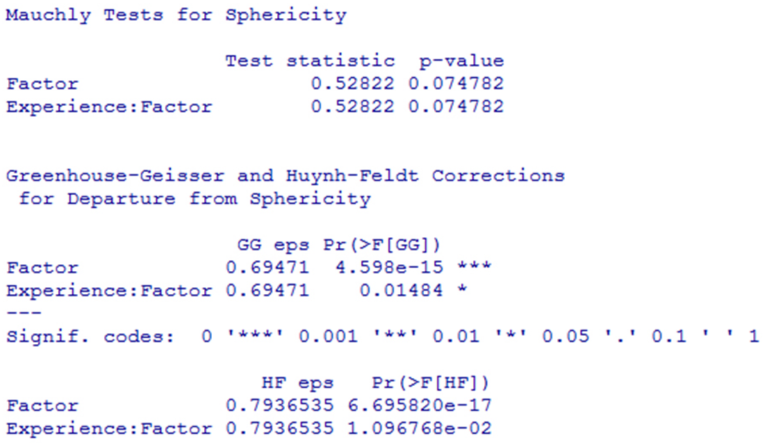
Figure IX.7. Panel showing the degree of accuracy of the tasters along the time (left panel), the average value of the tasters with and without experience(central panel) and the interaction graph (right panel).





```
[1] "ANOVA with repeated and without repeated measures"

[[2]]

Univariate Type II Repeated-Measures ANOVA Assuming Sphericity

                       SS num Df Error SS den Df         F    Pr(>F)
(Intercept)        3374.2      1   54.896     17 1044.9128 < 2.2e-16 ***
Experience          183.6      1   54.896     17   56.8426 8.105e-07 ***
Factor              245.2      3   44.612     51   93.4530 < 2.2e-16 ***
Experience:Factor    12.2      3   44.612     51    4.6679  0.005876 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A requirement of the statistical *F* in the ANOVA with repeated measures is that the variances of the differences between each two levels of the intra-subject factor, time, must be equal, and the levels independent of each other. This assumption is equivalent to stating that the variance-covariance matrix is proportional to the unit matrix (equal variances and null covariances),or that the joint dispersion is spherical. The table shows the results of the test of sphericity of Mauchley. As the value *p* is greater than 0.05 (p = 0.075) the assumption of sphericity is accepted.



```
Mauchly Tests for Sphericity

                  Test statistic  p-value
Factor                   0.52822 0.074782
Experience:Factor        0.52822 0.074782


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

                   GG eps Pr(>F[GG])
Factor            0.69471  4.598e-15 ***
Experience:Factor 0.69471    0.01484 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                     HF eps    Pr(>F[HF])
Factor            0.7936535 6.695820e-17
Experience:Factor 0.7936535 1.096768e-02
```

If the requirement of sphericity might not be fulfilled (if the value*p* is less than 0.05), one of the two modifications that are also shown should be used: Greenhouse-Geisser and Huynh-Feldt, to correct the degrees of freedom downward which allows to adjust (increase) the *p* value of the ANOVA contrast. In this case both corrected *p* values $4.598 * 10^{-15}$ and $6.69 * 10^{-17}$ are virtually zero, so it also shows a significant relationship between time and accuracy.

Value

A TXT file with the results of the ANOVA is obtained, and a boxplot or a beanplot can be displayed for each factor, in addition to showing an interaction graph of each of the factors with the time factor.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) Tratamiento de datos con R, STATISTICA y SPSS. Ediciones Díaz de Santos, Madrid, 978 pp.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Examples

```
## Not run:

data(ZIX5)

#Variation in the ability to ascertain the origin of
#a wine versus time and sommelier experience
```

```
IX6(data = ZIX5, varTime = c("Hour","Day","Week","Month"), Factor1 = "Experience",
Error = "Sommelier", mfrow = c(1,3), YLAB = "Degree of success",
PART = c("cex.axis = 1.5", "cex.lab = 1.8", "font.lab = 2", "mar = c(5,5,3,2)"),
PAR1 = c("cex.axis = 1.5", "cex.lab = 1.8", "font.lab = 2", "mar = c(5,5,3,2)"),
 INTERPLOT1 = c("ylab=YLAB", "xlab=XLAB", "legend=TRUE", "main='Factor Experience'",
"cex.main=2", "cex.lab=1.8", "col=rainbow(length(unique(datos4[,Factor1])))",
"trace.label=''"))

## End(Not run)
```

IX7 *ANCOVA FOR FACTORS WITH REPEATED AND NON-REPEATED MEASURES*

Description

An ANCOVA and a covariate in which there are factors with repeated measures (factors intra-subject) and factors without repeated measures (factors inter-subject) are applied. In addition to displaying the data with a boxplot or beanplot and an interaction graph of each of the factors along with the time factor.

Usage

```
IX7(data, varTime, Factor1, Covariable, Error, Factor2=NULL, Factor3=NULL, SS=2,
ResetPAR=TRUE, mfrow=NULL, YLAB="Dependent variable",
XLAB="Time", graphT="Boxplot", PART=NULL, LabelCatT=NULL,
COLORT=NULL, BOXPLOTT=NULL, BEANPLOTT=NULL, LEGENDT=NULL, AXIST=NULL,
MTEXTT= NULL, TEXTT=NULL, graph1="Boxplot", PAR1=NULL, OrderCat1=NULL,
LabelCat1=NULL, COLOR1=NULL, BOXPLOT1=NULL, BEANPLOT1=NULL, LEGEND1=NULL,
AXIS1=NULL, MTEXT1= NULL, TEXT1=NULL, graph2="Boxplot",  PAR2=NULL,
OrderCat2=NULL, LabelCat2=NULL, COLOR2=NULL, BOXPLOT2=NULL, BEANPLOT2=NULL,
LEGEND2=NULL, AXIS2=NULL, MTEXT2= NULL, TEXT2=NULL, graph3="Boxplot",
PAR3=NULL, OrderCat3=NULL, LabelCat3=NULL, COLOR3=NULL, BOXPLOT3=NULL,
BEANPLOT3=NULL, LEGEND3=NULL, AXIS3=NULL, MTEXT3= NULL, TEXT3=NULL,
INTERPLOT1=NULL, INTERPLOT2=NULL, INTERPLOT3=NULL, file="Output.txt")
```

Arguments

`data`	Data file.
`varTime`	Columns in which the dependent variable has been measured over time.
`Factor1`	First factor.
`Covariable`	Covariate.
`Error`	It is the variable that encodes the subjects studied.
`Factor2`	Second factor.
`Factor3`	Third factor.

SS When there are several factors, the decomposition of the sum of squares in parts attributed to each one of the factors is not unique. Four types of sums of squares are commonly used: type I for nested models, type II for balanced models, type III for unbalanced models (also balanced), and type IV for models with empty boxes. As a model of repeated measures all the boxes of the model have equal number of data (balanced model), the type II will be used by default. This description is very simplified; interested readers should consult supplementary bibliography for a correct use of the ANOVA with complex models.

ResetPAR If FALSE, the conditions are not placed by default in the function PAR and those defined by the user on previous graphics are maintained.

mfrow If it is NULL and there are various graphics, these appear in separate windows. If the graphics go into panels, this argument is a vector with the format c(nr, nc) indicating the number of figures per row (nr) and column (nc), by first filling the rows.

YLAB A legend for the Y axis in the boxplot and beanplot.

XLAB A legend for the X axis in the boxplot and beanplot.

graphT If NULL, the graphic of factor time is not performed and the other options are "Boxplot" or "Beanplot".

PART This accesses the function PAR which allows to modify many different aspects of the graph of the time factor.

LabelCatT This allows to specify a vector with the names of the categories of the graph of the time factor.

COLORT Vector with the color of the categories of the graph of the time factor.

BOXPLOTT It allows to specify the characteristics of the boxplot graph of the time factor.

BEANPLOTT It allows to specify the characteristics of the beanplot graph of the time factor.

LEGENDT It allows to include a legend to the graph of the time factor.

AXIST It allows to add axes to the graph of the time factor.

MTEXTT It allows to add text in the margins of the graph of the time factor.

TEXTT It allows to add text in any area of the inner part of the graph of the time factor.

graph1 If NULL, the graphic of factor 1 is not performed and the other options are "Boxplot" or "Beanplot".

PAR1 This accesses the function PAR which allows to modify many different aspects of the graph of the factor 1.

OrderCat1 It allows to specify a vector with the order in which the categories are shown in graph 1.

LabelCat1 It allows to specify a vector with the names of the categories of graph 1.

COLOR1 Vector with the color of the categories of graph 1.

BOXPLOT1 It allows to specify the characteristics of the boxplot of factor 1.

BEANPLOT1 It allows to specify the characteristics of the beanplot of factor 1.

LEGEND1 It allows to include a legend to the graph of factor 1.

AXIS1 It allows to add axes to the graph of factor 1.

`MTEXT1` It allows to add texts on the margins of the graph of factor 1.

`TEXT1` It allows to add text in any area of the inner part of the graph of factor 1.

`graph2` If NULL, the graphic of factor 2 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR2` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 2.

`OrderCat2` It allows to specify a vector with the order in which the categories are shown in graph 2.

`LabelCat2` It allows to specify a vector with the names of the categories of graph 2.

`COLOR2` Vector with the color of the categories of graph 2.

`BOXPLOT2` It allows to specify the characteristics of the boxplot of factor 2.

`BEANPLOT2` It allows to specify the characteristics of the beanplot of factor 2.

`LEGEND2` It allows to include a legend to the graph of the factor 2.

`AXIS2` It allows to add axes to the graph of the factor 2.

`MTEXT2` It allows to add text in the margins of the graph of the factor 2.

`TEXT2` It allows to add text in any area of the inner part of the graph of the factor 2.

`graph3` If NULL, the graphic of factor 3 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR3` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 3.

`OrderCat3` It allows to specify a vector with the order in which the categories are shown in graph 3.

`LabelCat3` It allows to specify a vector with the names of the categories of graph 3.

`COLOR3` Vector with the color of the categories of graph 3.

`BOXPLOT3` It allows to specify the characteristics of the boxplot of factor 3.

`BEANPLOT3` It allows to specify the characteristics of the beanplot of factor 3.

`LEGEND3` It allows to include a legend to the graph of the factor 3.

`AXIS3` It allows to add axes to the graph of the factor 3.

`MTEXT3` It allows to add text in the margins of the graph of the factor 3.

`TEXT3` It allows to add text in any area of the inner part of the graph of the factor 3.

`INTERPLOT1` It allows to specify the characteristics of the graph of interaction of factor 1.

`INTERPLOT2` It allows to specify the characteristics of the graph of interaction of factor 2.

`INTERPLOT3` It allows to specify the characteristics of the graph of interaction of factor 3.

`file` TXT FILE. Output file name.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.1.7. ANOVA for factors with repeated and non-repeated measures

See section *details* of the function IX5 or Guisande et al. (2011) for an explanation on this type of ANOVA.

FUNCTIONS

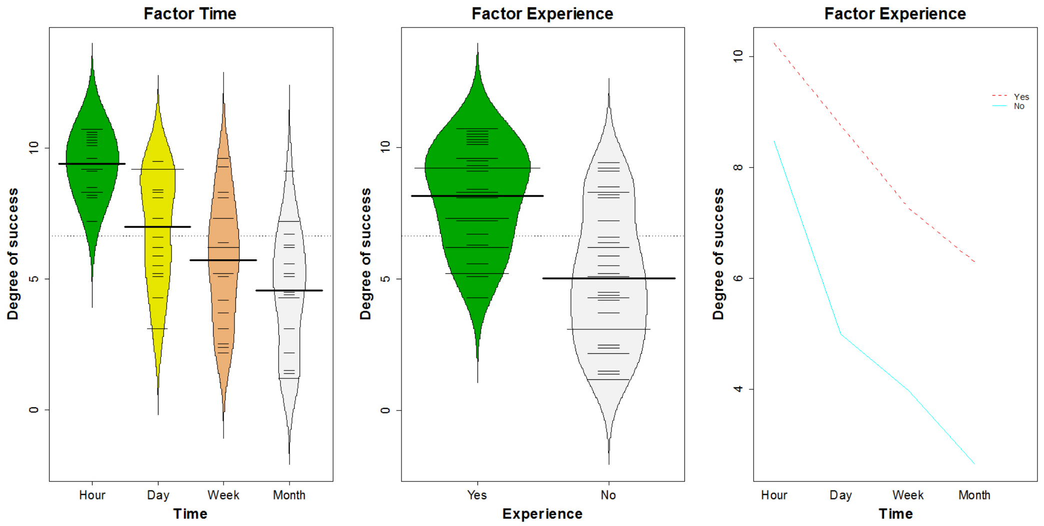
For the beanplot, the function beanplot of the package beanplot (Kampstra, 2008; Kampstra, 2013)is used. The linear model is performed with the functionlm, the ANOVA with the function aov and the graph of interaction is performed with the function interaction.plot, all of the stats package.

EXAMPLE

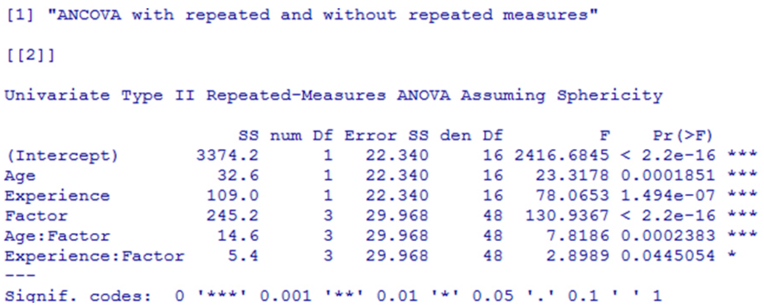
In an experiment conducted with expert tasters and people who had no experience of tasting, they were taught to identify 15 types of wines from different regions. The objective was to determine whether the ability to ascertain the provenance of the wine varied over time (after one hour, one day, one week and one month), and depending on the experience. To do this, every time each person assessed a large number of samples and the degree of success on a scale of 0 to 12 was recorded. In addition, the possible influence of the age of the tasters, which was considered as a covariate was taken into account.

Figure IX.8 clearly shows how the degree of accuracy of the tasters decreases over time (left panel), but experienced tasters have a greater degree of success than those who do not have (central panel) and finally the interaction graph seems to show that the degree of success, or intensity of memory, decreases over time from one hour to one month, similar in both groups (both lines are approximately parallel, especially between one day and one month). The intensity is smaller as time passes, and experts always maintain a degree of success two or three points higher than non-experts. However, it is necessary to check whether this difference is statistically significant.

Figure IX.8. Panel showing the degree of accuracy of the tasters along the time (left panel), the average value of the tasters with and without experience(central panel) and the interaction graph (right panel).

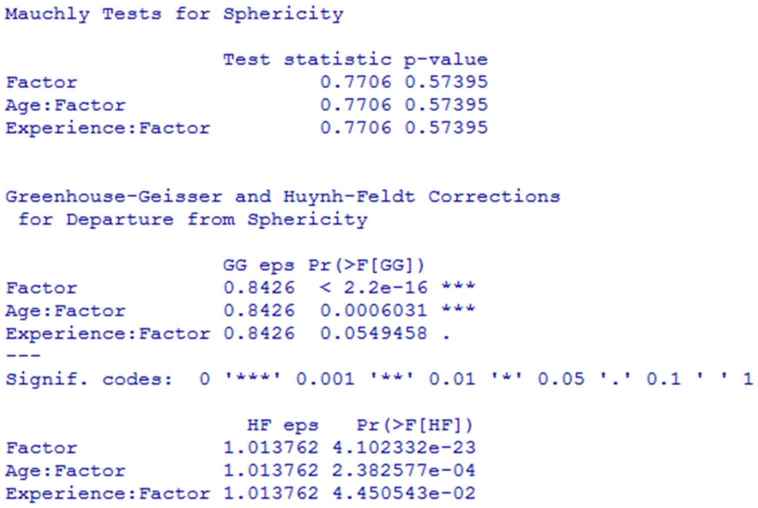


The «Experience» factor is significant, with $p < 0.001$. The factor «Time» is also significant($p < 0.001$), as we had already checked, and the interaction between them, with $p = 0.044$, which means that the evolution of the memory is different in both groups, and the curves are not really parallel as it had seemed before. Therefore, the ability to identify the origin of wine decreases with different intensity over time between experts and non-experts. Finally, the age also has a significant effect on the degree of success, with $p < 0.001$.



```
[1] "ANCOVA with repeated and without repeated measures"

[[2]]

Univariate Type II Repeated-Measures ANOVA Assuming Sphericity

                      SS num Df Error SS den Df         F    Pr(>F)
(Intercept)       3374.2      1   22.340     16 2416.6845 < 2.2e-16 ***
Age                 32.6      1   22.340     16   23.3178 0.0001851 ***
Experience         109.0      1   22.340     16   78.0653 1.494e-07 ***
Factor             245.2      3   29.968     48  130.9367 < 2.2e-16 ***
Age:Factor          14.6      3   29.968     48    7.8186 0.0002383 ***
Experience:Factor    5.4      3   29.968     48    2.8989 0.0445054 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A requirement of the statistical *F* in the ANOVA with repeated measures is that the variances of the differences between each two levels of the intra-subject factor, time, must be equal, and the levels independent of each other. This assumption is equivalent to stating that the variance-covariance matrix is proportional to the unit matrix (equal variances and null covariances),or that the joint dispersion is spherical. The table shows the results of the test of sphericity of Mauchley. As the value *p* is greater than 0.05 ($p = 0.57$) the assumption of sphericity is accepted.



```
Mauchly Tests for Sphericity

                  Test statistic p-value
Factor                    0.7706 0.57395
Age:Factor                0.7706 0.57395
Experience:Factor         0.7706 0.57395


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

                  GG eps Pr(>F[GG])
Factor            0.8426  < 2.2e-16 ***
Age:Factor        0.8426  0.0006031 ***
Experience:Factor 0.8426  0.0549458 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                    HF eps    Pr(>F[HF])
Factor            1.013762 4.102332e-23
Age:Factor        1.013762 2.382577e-04
Experience:Factor 1.013762 4.450543e-02
```

If the requirement of sphericity might not be fulfilled (if the value*p* is less than 0.05), one of the two modifications that are also shown should be used: Greenhouse-Geisser and Huynh-Feldt, to correct the degrees of freedom downward which allows to adjust (increase) the *p value* of the ANOVA contrast. In this case both corrected *p* values $2.2 * 10^{-16}$ and $4.1 * 10^{-23}$ are virtually zero, so it also shows a significant relationship between time and accuracy.

Value

A TXT file with the results of the ANOVA is obtained, and a boxplot or a beanplot can be displayed for each factor, in addition to showing an interaction graph of each of the factors with the time factor.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) Tratamiento de datos con R, STATISTICA y SPSS. Ediciones Díaz de Santos, Madrid, 978 pp.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Examples

```
## Not run:

data(ZIX5)

#Variation in the ability to ascertain the origin of
#a wine versus time and, age and experience of the sommelier

IX7(data = ZIX5, varTime = c("Hour","Day","Week","Month"), Factor1 = "Experience",
Covariable = "Age" , Error = "Sommelier", mfrow = c(1,3), YLAB = "Degree of success",
graphT = "Beanplot", graph1 = "Beanplot", PART = c("cex.axis = 1.5", "cex.lab = 1.8",
"font.lab = 2", "mar = c(5,5,3,2)"), PAR1 = c("cex.axis = 1.5", "cex.lab = 1.8",
"font.lab = 2", "mar = c(5,5,3,2)"), INTERPLOT1 = c("ylab=YLAB", "xlab=XLAB",
"legend=TRUE", "main='Factor Experience'","cex.main=2", "cex.lab=1.8",
"col=rainbow(length(unique(datos4[,Factor1])))", "trace.label=''"))

## End(Not run)
```

IX8 *NESTED ANOVA*

Description

A nested analysis of variance is applied and the data can be displayed with a boxplot or a beanplot.

Usage

```
IX8(data, variables, Factor1, CF1=NULL, Factor2, CF2=1, Factor3=NULL,
CF3=NULL, Factor4=NULL, CF4=NULL, combine=FALSE,  trans=NULL, ResetPAR=TRUE,
mfrow=NULL, graph1="Boxplot", PAR1=NULL,  OrderCat1=NULL, LabelCat1=NULL,
COLOR1=NULL, BOXPLOT1=NULL, BEANPLOT1=NULL, LEGEND1=NULL, AXIS1=NULL,
MTEXT1= NULL, TEXT1=NULL, graph2="Boxplot", PAR2=NULL, OrderCat2=NULL,
LabelCat2=NULL, COLOR2=NULL, BOXPLOT2=NULL, BEANPLOT2=NULL, LEGEND2=NULL,
AXIS2=NULL, MTEXT2= NULL, TEXT2=NULL, graph3="Boxplot", PAR3=NULL,
OrderCat3=NULL, LabelCat3=NULL, COLOR3=NULL, BOXPLOT3=NULL, BEANPLOT3=NULL,
LEGEND3=NULL, AXIS3=NULL, MTEXT3= NULL, TEXT3=NULL, graph4="Boxplot",
PAR4=NULL, OrderCat4=NULL, LabelCat4=NULL, COLOR4=NULL, BOXPLOT4=NULL,
BEANPLOT4=NULL, LEGEND4=NULL, AXIS4=NULL, MTEXT4= NULL, TEXT4=NULL,
file="Output.txt")
```

Arguments

`data` Data file.

`variables` Dependent variable.

`Factor1` First factor.

`CF1` Factors of which the factor1 nests. If this factor does not nest any of them, it becomes NULL, which is the default option.

`Factor2` Second factor.

`CF2` Factors of which the factor2 nests. For example, a value of 1 would mean that nests the factor 1.

`Factor3` Third factor.

`CF3` Factors of which the factor3 nests. For example, a vector c(2,1) would mean that nests the factors 2 and 1. A value of 2 means that nests only the factor 2.

`Factor4` Fourth factor.

`CF4` Factors of which the factor4 nests. For example, a vector c(3,2,1) would mean that nests the factors 3, 2 and 1. A vector c(3,1) would mean that nests the factors 3 and 1.

`combine` If TRUE, a graphic panel that combines factors is shown. However, if the number of combinations is greater than 12, then it automatically turns itself off and becomes FALSE.

`trans` Type of transformation that is applied to the data:

1. NULL (Untransformed)
2. 1/x2
3. 1/x
4. LN
5. LOG
6. SQR (square root)
7. x2
8. x3
9. EXP (exponential)
10. ASN (arcsine)

`ResetPAR` If FALSE, the conditions are not placed by default in the function PAR and those defined by the user on previous graphs are maintained.

`mfrow` If it is NULL and there are various graphics, these appear in separate windows. If the graphics go into panels, this argument is a vector with the format c(nr, nc) indicating the number of figures per row (nr) and column (nc), by first filling the rows.

`graph1` If NULL, the graphic of factor 1 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR1` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 1.

`OrderCat1` It allows to specify a vector with the order in which the categories of the graph 1 are shown.

`LabelCat1` It allows to specify a vector with the names of the categories of the graph of the factor 1.

`COLOR1` A vector with the color of the categories of graph 1.

`BOXPLOT1` It allows to specify the characteristics of the boxplot of the factor 1.

`BEANPLOT1` It allows to specify the characteristics of the beanplot of the factor 1.

`LEGEND1` It allows to include a legend to the graph of the factor 1.

`AXIS1` It allows to add axes to the graph of the factor 1.

`MTEXT1` It allows to add text in the margins of the factor 1.

`TEXT1` It allows to add text in any area of the inner part of the graph of the factor 1.

`graph2` If NULL, the graphic of factor 2 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR2` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 2.

`OrderCat2` It allows to specify a vector with the order in which the categories of the graph 2 are shown.

`LabelCat2` It allows to specify a vector with the names of the categories of the graph 2.

`COLOR2` Vector with the color of the categories of graph 2.

`BOXPLOT2` It allows to specify the characteristics of the boxplot of the factor 2.

`BEANPLOT2` It allows to specify the characteristics of the beanplot of the factor 2.

`LEGEND2` It allows to include a legend to the graph of the factor 2.

`AXIS2` It allows to add axes to the graph of the factor 2.

`MTEXT2` It allows to add text in the margins of the factor 2.

`TEXT2` It allows to add text in any area of the inner part of the graph of the factor 2.

`graph3` If NULL, the graphic of factor 3 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR3` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 3.

`OrderCat3` It allows to specify a vector with the order in which the categories of the graph 3 are shown.

`LabelCat3` It allows to specify a vector with the names of the categories of the graph 3.

`COLOR3` Vector with the color of the categories of graph 3.

`BOXPLOT3` It allows to specify the characteristics of the boxplot of the factor 3.

`BEANPLOT3` It allows to specify the characteristics of the beanplot of the factor 3.

`LEGEND3` It allows to include a legend to the graph of the factor 3.

`AXIS3` It allows to add axes to the graph of the factor 3.

`MTEXT3` It allows to add text in the margins of the factor 3.

`TEXT3` It allows to add text in any area of the inner part of the graph of the factor 3.

`graph4` If NULL, the graphic of factor 4 is not performed and the other options are "Boxplot" or "Beanplot".

`PAR4` This accesses the function PAR which allows to modify many different aspects of the graph of the factor 4.

`OrderCat4` It allows to specify a vector with the order in which the categories of the graph 4 are shown.

`LabelCat4` It allows to specify a vector with the names of the categories of the graph 4.

`COLOR4` Vector with the color of the categories of graph 4.

`BOXPLOT4` It allows to specify the characteristics of the boxplot of the factor 4.

`BEANPLOT4` It allows to specify the characteristics of the beanplot of the factor 4.

`LEGEND4` It allows to include a legend to the graph of the factor 4.

`AXIS4` It allows to add axes to the graph of the factor 4.

`MTEXT4` It allows to add text in the margins of the factor 4.

`TEXT4` It allows to add text in any area of the inner part of the graph of the factor 4.

`file` TXT FILE. Output file name with the results.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.1.8. Nested ANOVA

The nested ANOVA is an experimental design in which the levels of one or more factors are dependent on the levels of other factors. It is an extension of a one-way ANOVA, which divides each group into subgroups. It is said that a factor B is nested in another factor (or that their levels are nested in the (A) when all levels of the factor B are associated with a single level of factor A. This is denoted as B included in A. It is possible to have multiple nested factors in the experimental design, i.e., having groups, subgroups for each group, sub-subgroups for each subgroup, etc.

Nested Analysis of Variance has a null hypothesis for each factor. In a two-way nested ANOVA, a null hypothesis would be that the subgroups within each group have the same media, the second null hypothesis would be that groups have the same mean. Nested ANOVA requirements are equal to those of the ANOVA, i.e., it is required that waste will have Normal distribution and homogeneity of variances.

An example of a nested ANOVA is planting a tree species in three different habitats, for example, in the valley, on the slope of the mountain and in the high mountain, and repeat this experimental design in two different areas, for example, an area that is a natural park and, therefore, there is a degree of protection, and another area which does not have any type of control. The objective would be to determine if there are variations in the rate of tree growth (dependent variable) between the two areas (primary factor), and between the habitats within each zone (nested secondary factor).

In another example, it is assumed that you are testing the null hypothesis if monkeys under stress and not subjected to stress have the same amount of antibodies. As the monkeys occupy much space, it is necessary to take them in four different rooms and in each one of them there are monkeys of the two types, those who have been subjected and has not subjected to stress.

It may be the case that the conditions of the rooms (light, temperature, etc.) can cause some differences in the level of stress and, therefore, perhaps in the level of antibodies. To eliminate this potential effect of the different rooms in the experimental design, a nested ANOVA in which the first factor would be the rooms (with four levels, the 4 rooms) and nested factor would be the level of stress of the monkeys (with two levels, stress and not stressed) should be performed. With this design it can be observed if there are differences in the amount of antibodies (dependent variable) between rooms and in the amount of antibodies in the group of stressed and unstressed monkeys in each room.

In addition to reporting any differences in means between the levels of each factor, in means between each subfactor level within each factor, etc., the nested ANOVA is very useful for future experimental designs because it indicates the degree of variability within each factor.

For example, imagine that the levels of glycogen in various types of muscles in rats subjected and not subjected to stress are measured. Therefore, the muscle factor would be nested within the stress factor and the dependent variable would be the level of glycogen. It may be the case that the variability between muscle glycogen content is very small, while among rats is high. Therefore, in the next design, the effort should be focused on measuring glycogen in a greater number of rats and only in a muscle within each rat, that is to say, it would no longer be necessary a nested ANOVA as this would be a single factor, the level of stress to which the rats are subjected.

FUNCTIONS

The function lillie.test of the package nortest (Gross, 2013) is used to perform the Normality test of Kolmogorov-Smirnov with Lilliefors' correction.

The function shapiro.test of the package base stats is used to perform the Shapiro-Wilk test. The ANOVA is performed with the function Anova of the package car (Fox et al., 2014). For the beanplot the function beanplot of the package beanplot (Kampstra, 2008; Kampstra, 2013)is used.

The Levene test is performed with the function levene.test of the packagelawstat (Gastwirth et al., 2013). The asymmetry and the kurtosis are carried out with the functions skewness and kurtosis, respectively, of the package e1071 (Meyer et al., 2014).

EXAMPLE 1

Step 1.

Lichens are organisms considered as bioindicators of pollution, since it has been demonstrated that their presence is often lower in most polluted areas.

The data used in the example is the average of lichen species found in two species of tree, *Populus alba* (White poplar) and *Aesculus hippocastanum* (Buckeye), in three cities in Spain (Madrid, Barcelona and Seville), in three types of environments (low, medium and high levels of air pollution) and also showing the researcher who performed the counting of species, by whether this data could be affected depending on the person performing the sampling. The average is the result of counting the lichen species in 40 trees of each species (*Populus alba* and *Aesculus hippocastanum*), in each of the cities, in the three types of environments and and for each of the researchers.

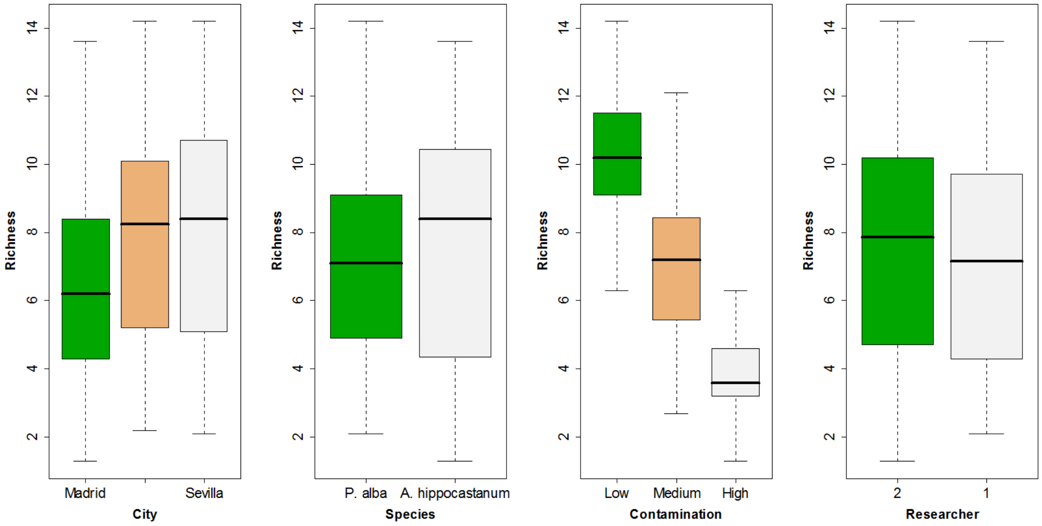
The objective is to determine if the number of lichen species varies depending on the researcher, who recounts. As the number of species of lichens may be affected by the species of tree and/or the degree of pollution of the cities, it is necessary to perform a nested ANOVA.

The most important thing in a nested ANOVA is to nest the factors in a correct way. In this study, the factor 1 (City) is not nest and, therefore, the default conditions are left *CF1=NULL*. Factor 2

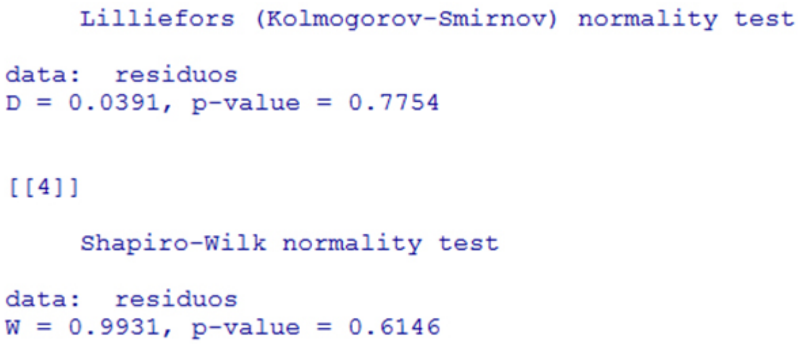
(Species) is nested in 1 *CF2=1*. Factor 3 is nested in factors 2 and 1 *CF3=c(2,1)*. Finally, the researcher factor is nested in the previous three *CF4=c(3,2,1)*.

Figure IX.9 shows the mean and range of the richness of lichens for each of the categories of the factors analyzed. It is noted that richness is higher in Barcelona and Seville, and lower in Madrid. There is more richness in the species *Aesculus hippocastanum* than in *Populus alba*, although it is necessary to check if the differences are significant. It is clear that the richness decreases with increasing pollution and, in principle, the two researchers made a recount of similar species of lichens.

Figure IX.9. Panel showing the average and range of the richness of lichens between cities, species, degree of contamination and between researchers.

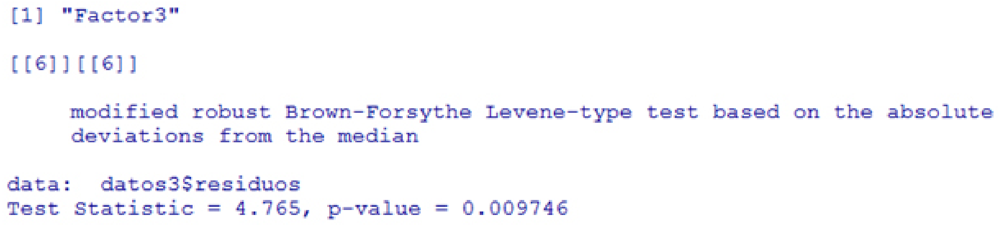


First, the normality of the residuals is taken into account. It is noted that both the Kolmogorov-Smirnov test with the correction of Lilliefors (p = 0.775), as the Shapiro-Wilk (p = 0.615) show p > 0.05 and, therefore, the residuals comply with a Normal distribution.



```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0391, p-value = 0.7754


[[4]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9931, p-value = 0.6146
```

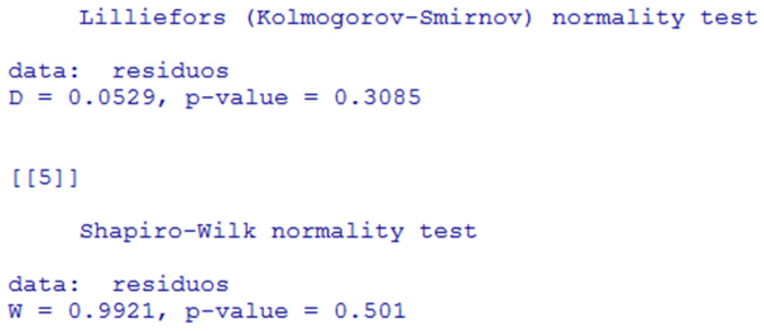
The Levene test considering the median shows that it also meets the homogeneity of variances (p > 0.05) for almost factors, with the exception of the factor 3 (contamination). Therefore, it is necessary to perform a transformation.



```
[1] "Factor3"

[[6]][[6]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 4.765, p-value = 0.009746
```

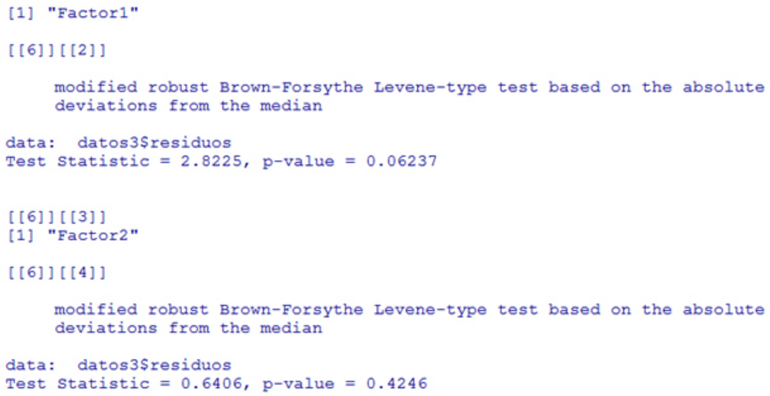
Step 2.

The data is transformed using the square root of the argument *trans = "SQR"*. It is noted that both the Kolmogorov-Smirnov test with the correction of Lilliefors (p = 0.308), as the Shapiro-Wilk (p = 0.501) show p > 0.05 and, therefore, the residuals comply with a Normal distribution.

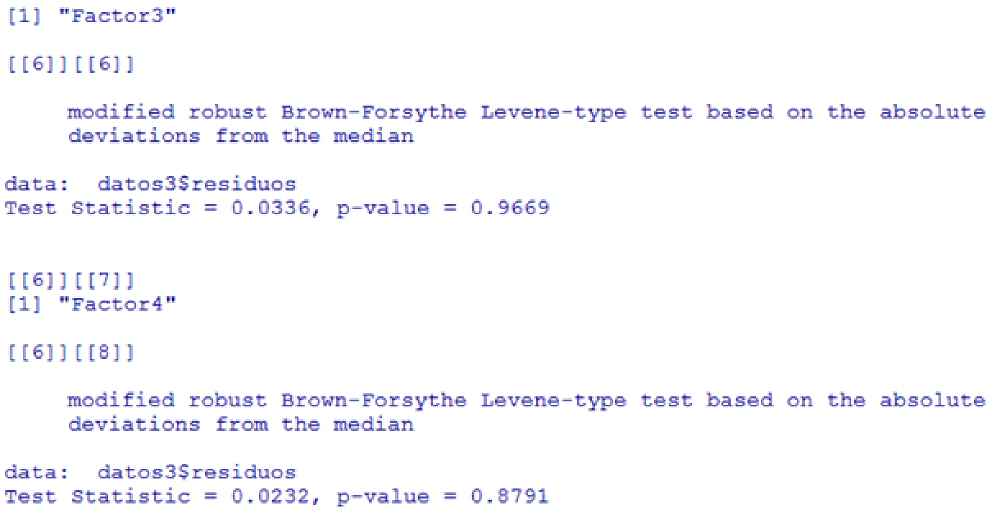


```
     Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0529, p-value = 0.3085


[[5]]

     Shapiro-Wilk normality test

data:  residuos
W = 0.9921, p-value = 0.501
```

In addition, there is homogeneity of variances in all factors, as p > 0.05.

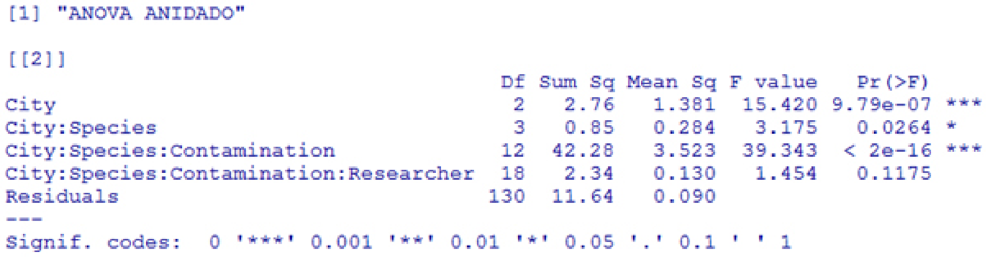


```
[1] "Factor1"

[[6]][[2]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 2.8225, p-value = 0.06237


[[6]][[3]]
[1] "Factor2"

[[6]][[4]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 0.6406, p-value = 0.4246
```



```
[1] "Factor3"

[[6]][[6]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 0.0336, p-value = 0.9669


[[6]][[7]]
[1] "Factor4"

[[6]][[8]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 0.0232, p-value = 0.8791
```

The ANOVA results are shown in the following table. There are significant differences between the cities in the number of species of lichens (ANOVA, $F_{2,130}$= 15.42, p < 0.001), and also between the two species of trees (ANOVA, $F_{3,130}$= 3.17, p = 0.026). There are clear differences in the number of species of lichens observed depending on the degree of contamination (ANOVA, $F_{12,130}$= 39.343, p < 0.001). In regard to the purpose of the example, the two researchers conducted a count of lichens such as there are no significant differences between them (ANOVA, $F_{18,30}$= 1.45, p = 0.117).



```
[1] "ANOVA ANIDADO"

[[2]]
                                         Df Sum Sq Mean Sq F value   Pr(>F)    
City                                      2   2.76   1.381  15.420 9.79e-07 ***
City:Species                              3   0.85   0.284   3.175   0.0264 *  
City:Species:Contamination               12  42.28   3.523  39.343  < 2e-16 ***
City:Species:Contamination:Researcher    18   2.34   0.130   1.454   0.1175    
Residuals                               130  11.64   0.090                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

EXAMPLE 2

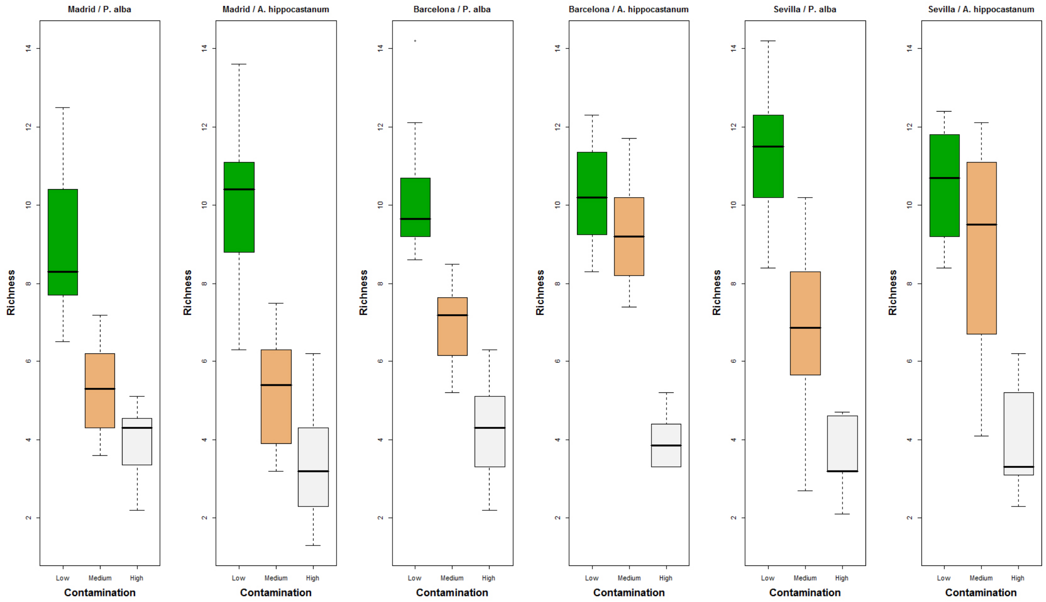
In the second example this only works with three factors to determine if the wealth of lichens varies depending on the degree of pollution, thus excluding research factor.

In addition there is another nesting assumption, in such a way that the factor 2 (species) is nestled in 1 *CF2=1*, but factor 3 only nests the 2 and not the 1 *CF3=2*.

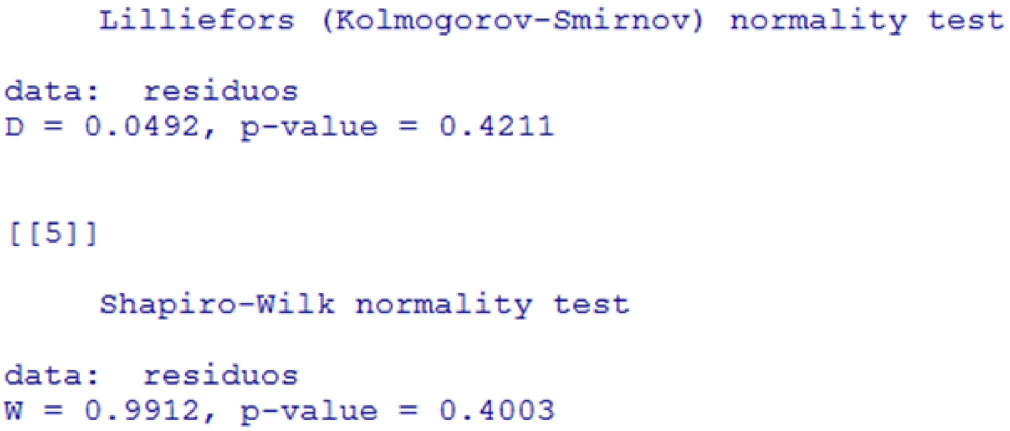
The argument *combine=TRUE*, which means that a graphical panel that combines the factors (Figure IX.10) is carried out automatically. As the third factor is the pollution, which is shown on the X axis of all graphs and combined with each of the categories of the other factors.

As the third factor which is combined with other factors, all changes must be made at the level of this third factor. For example, changing the order of the categories with the argument *OrderCat3=c("Low","Medium","High")*, or modifications to the graph with the argument *PAR3*. It is possible that the error. *figure margins too large*, in which case the graph window is maximized and the script is run again.

Figure IX.10. Panel showing the average and range of the richness of lichens for each level of pollution combined with categories of other factors.

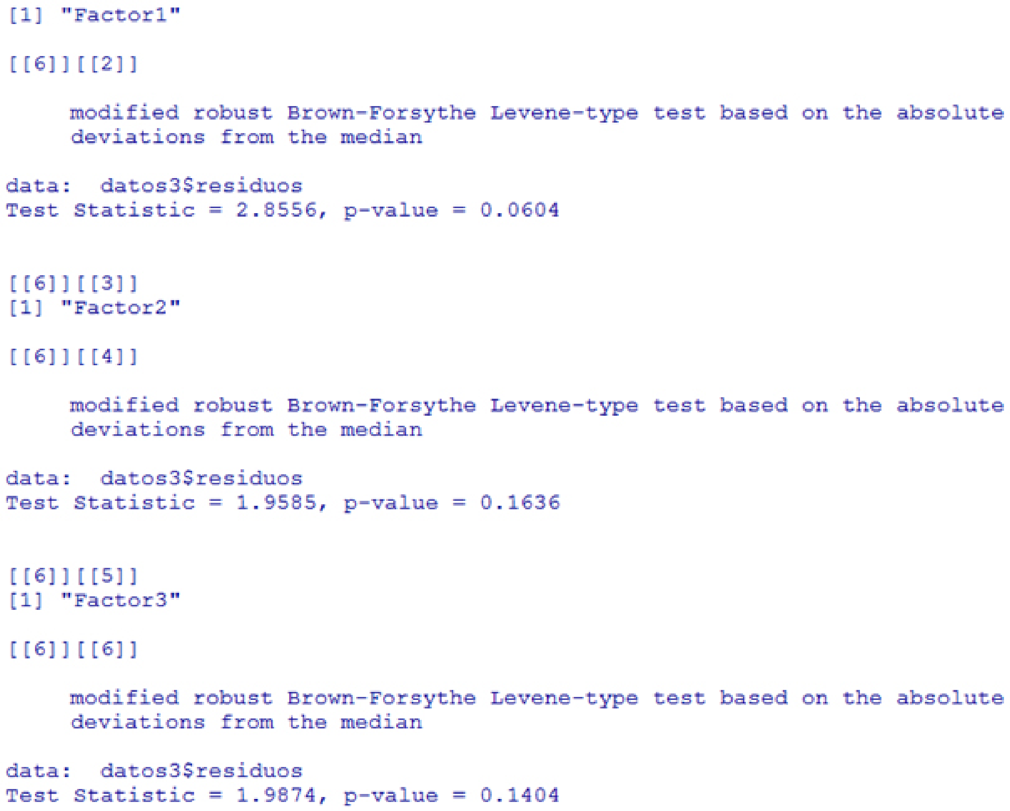


First we look at the normality of the residuals. It is noted that both the Kolmogorov-Smirnov test with Lilliefors correction (p = 0.421), as the Shapiro-Wilk (p = 0.4) shows p > 0.05 and, therefore, the residuals comply with a normal distribution.



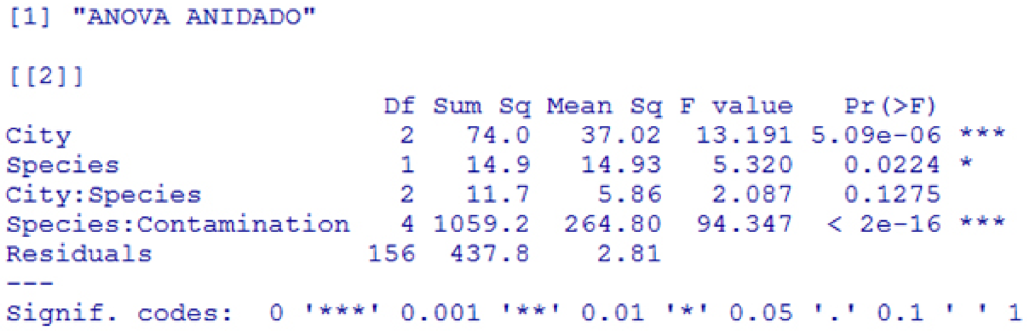
```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0492, p-value = 0.4211


[[5]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9912, p-value = 0.4003
```

The Levene test considering the median shows that also meets the homogeneity of variances (p > 0.05) for all factors. Therefore, it is not necessary to transform the data.



```
[1] "Factor1"

[[6]][[2]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 2.8556, p-value = 0.0604


[[6]][[3]]
[1] "Factor2"

[[6]][[4]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 1.9585, p-value = 0.1636


[[6]][[5]]
[1] "Factor3"

[[6]][[6]]

     modified robust Brown-Forsythe Levene-type test based on the absolute
     deviations from the median

data:  datos3$residuos
Test Statistic = 1.9874, p-value = 0.1404
```

The results of the ANOVA are shown in the following table. There are significant differences between the cities in the number of species of lichens (ANOVA, $F_{2,156}$= 13.19, p < 0.001), but not between the two species of trees considering the situation in which is nested in the factor city (ANOVA, $F_{2,156}$= 2.09, p = 0.127). In regard to the purpose of the example, there are significant differences in the observed number of species of lichens depending on the degree of contamination (ANOVA, $F_{4,156}$= 94.347, p < 0.001).

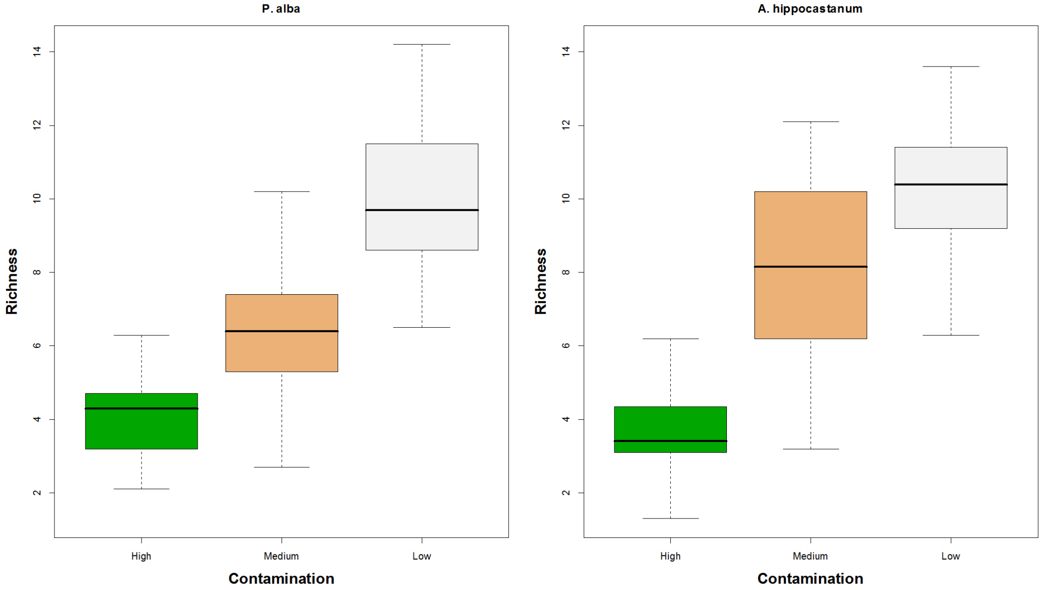


```
[1] "ANOVA ANIDADO"

[[2]]
                        Df Sum Sq Mean Sq F value   Pr(>F)    
City                     2   74.0   37.02  13.191 5.09e-06 ***
Species                  1   14.9   14.93   5.320   0.0224 *  
City:Species             2   11.7    5.86   2.087   0.1275    
Species:Contamination    4 1059.2  264.80  94.347  < 2e-16 ***
Residuals              156  437.8    2.81                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

EXAMPLE 3

Working with two factors in the last script(Figure IX.11).

Figure IX.11. Panel showing the average and range of the richness of lichens for each level of pollution combined with the species.



Value

A TXT file is obtained with the results of the nested ANOVA and a boxplot or beanplot can be represented for each factor or perform a combination of factors.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Gastwirth, J.L., Gel, Y.R., Hui, W.L.W., Lyubchich, V., Miao, W. & Noguchi, K. (2013). An R package for biostatistics, public policy, and law. R package version 2.4.1. Available at: http://CRAN.R-project.org/package=lawstat.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. & Lin, C.C (2014) Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. Available at: http://CRAN.R-project.org/package=e1071.

Examples

```
## Not run:

data(ZIX6)

#Case 1. Four factors

#Step 1

IX8(data = ZIX6 , variables = "Richness", Factor1 = "City", Factor2 = "Species",
CF2=1, Factor3 = "Contamination", CF3=c(2,1), Factor4 = "Researcher",
CF4=c(3,2,1), mfrow = c(1,4) , PAR1 = c("mar=c(5,5,3,2)", "font.lab=2",
"cex.lab=1.5","cex.axis=1.7"), OrderCat3 = c("Low","Medium","High"))


#Step 2. Data are transformed

IX8(data = ZIX6 , variables = "Richness", Factor1 = "City", Factor2 = "Species",
CF2=1, Factor3 = "Contamination", CF3=c(2,1), Factor4 = "Researcher",
CF4=c(3,2,1), mfrow = c(1,4), PAR1 = c("mar=c(5,5,3,2)", "font.lab=2",
"cex.lab=1.5","cex.axis=1.7"), OrderCat3 = c("Low","Medium","High"), trans = "SQR")

#Case 2.. Three factors (It is necessary a large monitor to
#view the plot)

windows(16,8)

IX8(data=ZIX6, variables="Richness",   Factor1 = "City", Factor2 = "Species",
CF2=1, Factor3 = "Contamination", CF3=2, combine=TRUE, OrderCat3=c("Low", "Medium","High"),
PAR3 = c("mar=c(5,5,3,2)","font.lab=2", "cex.lab=1.7") )

#Case 3. Two factors

IX8(data=ZIX6, variables="Richness", Factor1= "Species",
Factor2="Contamination",
CF2=1, combine=TRUE, OrderCat3=c("Low", "Medium","High"), PAR3 =c("mar=c(5,5,3,2)",
"font.lab=2", "cex.lab=1.7"))

## End(Not run)
```

IX9 *CONTRASTS OF NON-PARAMETRIC HOMOGENEITY FOR TWO INDEPENDENT SAMPLES*

Description

The Kolmogorov-Smirnov test for two samples and the sum of ranks of Wilcoxon are applied. In addition, the data can be represented with a boxplot or a beanplot.

Usage

```
IX9(data, variables, factor, pop1, pop2, graph="Boxplot", PAR=NULL,
ResetPAR=TRUE, YLAB=NULL, XLAB=NULL, OrderCat=NULL, LabelCat=NULL,
COLOR=NULL, BOXPLOT=NULL, BEANPLOT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL,
TEXT=NULL, file="Output.txt")
```

Arguments

`data`	Data file.
`variables`	Dependent variable.
`factor`	It defines which is the variable that acts as a factor.
`pop1`	First group of the population variable to compare.
`pop2`	Second group of the population variable to compare.
`graph`	If NULL, there is no graph and the other options are "Boxplot" or "Beanplot".
`PAR`	It accesses the function PAR which allows to modify many different aspects of the graph.
`ResetPAR`	If FALSE, the conditions are not placed by default in the PAR function and are those defined by the user in previous graphics.
`YLAB`	Legend of the Y axis in the boxplot and beanplot.
`XLAB`	Legend of the X axis in the boxplot and beanplot.
`OrderCat`	It allows to specify a vector with the order in which categories are displayed in the graph.
`LabelCat`	It allows to specify a vector with the names of the categories of the graph.
`COLOR`	Vector with the color of the categories of the graph.
`BOXPLOT`	It allows to specify the characteristics of the boxplot.
`BEANPLOT`	It allows to specify the characteristics of the beanplot.
`LEGEND`	It allows to include a legend.
`AXIS`	It allows to add axes.
`MTEXT`	It allows to add text in the margins of the graph.
`TEXT`	It allows to add text in any area of the inner part of the graph.
`file`	TXT FILE. Output file name with the results.

Details

IX. CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

IX.2. NON-PARAMETRIC TESTS

Often assumptions about normality and equality of variances in our data can not be assumed, therefore it is necessary to use the so-called non-parametric statistics or free distribution methods. Similarly, the non-parametric tests can be used when data of the population are in the form of range, when the procedure for obtaining the sample prevents assimilate it with a succession of quantitative values (Pérez, 2004).

They have the advantage that they are more generally applicable than the parametric ones (these contrasts can also be applied to normally distributed variables) since they do not require any conditions on the type of distribution. However, they are less sensitive to the detection of differences than parametric tests, although it can be said that the coincidence between the results obtained with the parametric and non-parametric tests is greater than 90%.

Within non-parametric tests of homogeneity, two different types can be distinguished depending on the nature of the samples. Independent samples would be obtained when the elements were randomly selected from a separate or non-related way (for example, a measure of the level of sugar in both men and women), and dependent samples, when the measurements are related, usually because they correspond to the same individual (for example, a measure of the sugar level in diabetics, before and after a treatment).

IX.2.1. Contrasts for two independent samples

The contrasts are similar to the parametric contrast *t* of Student. There are three main tests trying to determine if the differences observed between two samples are due to random or if they belong to two different populations. In all of them, the data is sorted in ascending without considering its membership group and, depending on its position, they are assigned an order number or range.

IX.2.1.1. Mann-Whitney U *test*

It is a test that compares the central tendency of two samples, that is not necessary to have the same size, on the basis of the null hypothesis that in both samples the central measure is the same.

It is a widely used contrast, although in the event that not only differences are measured in the central tendency of the data but on other features such as, for example, asymmetry or dispersion of data, it is more appropriate to use the test of Wald-Wolfowitz.

A premise of this test is that the values of each of the samples are different, i.e., that there is no overlap of data (repeated values). In case of overlapping, the contrast becomes more conservative (it is more difficult to find differences between samples). If the number of overlaps is very large, especially in small samples and with continuous variables, the use of the Kolmogorov-Smirnov test for two samples is advisable, although there is a correction for overlaps of the own statistician.

A range in ascending order (1,2,3, ...), is assigned to the ordered samples from lowest to highest, having as many ranges as data. In case of overlapping or draw, the ranges involved in it are added and divided by the number of overlapping data.

The complete development of this test is displayed in Sokal and Rohlf (1981). This is to calculate the statistic U with which is possible to know whether the mean of the ranks is significantly different between the two samples.

If the total number of data is less than 20, a table for determining critical values of U is used. However, if the total number of data is greater than 20, an approach to the Normal from statistical U can be used, which indicates the value Z of the Normal and its probability.

IX.2.1.2. Wald-Wolfowitz runs Test

It is a test that compares the distribution of two samples that do not need to have the same size, based on the null hypothesis that the distribution of ranges is random, i.e., the samples are homogeneous.

The data from the two samples are sorted together in ascending order and each set of values is assigned a run value. The streaks are sequences of values of the same group when the samples have been ordered.

A full description of the statistical appears in section 25.6 of the book *Biostatistical Analysis* (Zar, 1999).

If two samples have the same distribution it can be expected that in the sorting of data from least to greatest, both samples are very mixed (randomized), the number of runs is high.

With a contrast table is determined if the number of runs found is significantly large and the randomness of the ranges can be assumed and, therefore no differences between samples. If the number of data is large, as with the Mann-Whitney contrast, this statistic is set to Normal and the homogeneity of the samples can be analyzed by calculating the statistical *Z*.

This statistic is less powerful than the contrast of Mann-Whitney, but it has the advantage that detects differences not only of central tendency but dispersion and asymmetry.

IX.2.1.3. Kolmogorov-Smirnov Test for two samples

The advantage is that it measures the differences between the cumulative relative frequencies of the two samples, and that the differences are detected not only in the central tendency, but also in the dispersion and form (symmetry, pointing) of the samples.

When the sample size is large, the statistical U of Mann-Whitney is better. This statistical test is used to test that the H_0 of the two samples come from the same population, therefore this requires comparing two sampling distribution functions, observing the maximum difference between them.

The complete development of the calculation of this statistic is at Sokal & Rohlf (1981), in which the steps are:

1. Sorting the observations in each sample and distribute them in classes.

2. Determine the cumulative frequency of the classes in each sample.

3. Search for the maximum difference between the cumulative frequency of both samples.

4. Determine the critical values of D_{max} in a table.

5. If D_{max} is greater than the critical value, the null hypothesis that the samples come from the same population is rejected.

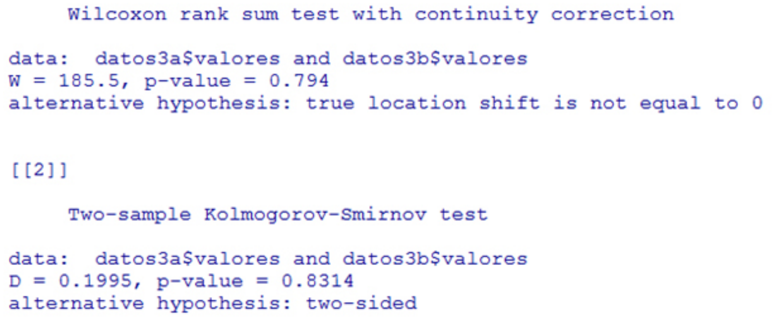
FUNCTIONS

For the beanplot, the function beanplot of the beanplot package (Kampstra, 2008; Kampstra, 2013) is used.

The Mann-Whitney U test is performed with the function wilcox.test and the Kolmogorov-Smirnovof test is performed for two samples with the function ks.test, both from the stats package.

EXAMPLE

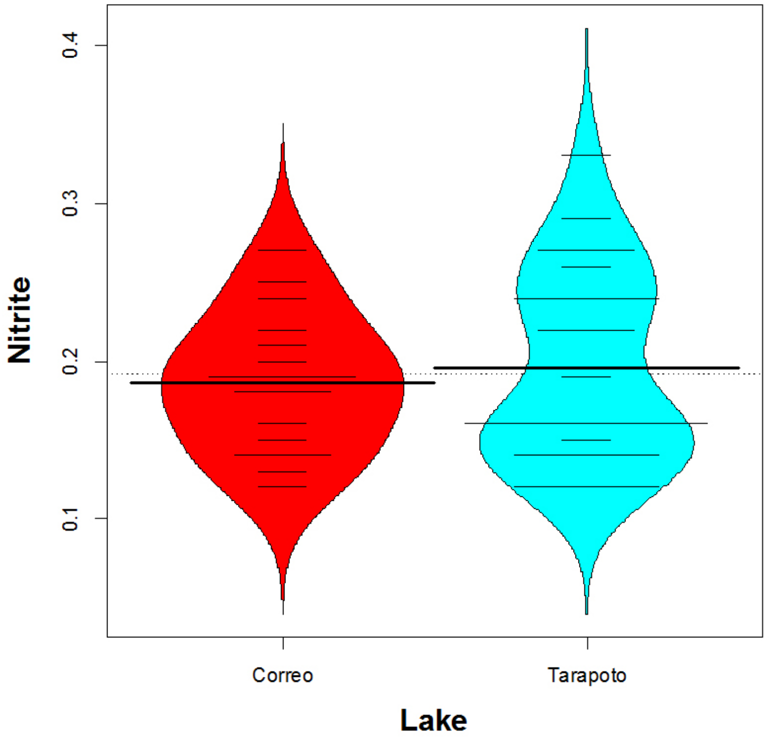
Data from the concentration of nitrite, nitrate and ammonium (μML) in the lakes of the Colombian Amazon in two different months. The aim is to determine whether there are differences in the concentration of nitrate among all lakes.



```
        Wilcoxon rank sum test with continuity correction

data:  datos3a$valores and datos3b$valores
W = 185.5, p-value = 0.794
alternative hypothesis: true location shift is not equal to 0


[[2]]

        Two-sample Kolmogorov-Smirnov test

data:  datos3a$valores and datos3b$valores
D = 0.1995, p-value = 0.8314
alternative hypothesis: two-sided
```

Figure IX.12 shows that the concentration is very similar in both lakes. The Wilcoxon rank-sum test (also known as U test of Mann-Whitney) with p = 0.794 accepts the hypothesis of equality, similar to the one of Kolmogorov-Smirnov test with p = 0.831. It is important to mention that when there are draws, that is to say, there are equal values in the groups to be compared, R does not calculate the exact value of probability, neither in the Wilcoxon sum of ranks nor in the Kolmogorov-Smirnov test and, therefore, small differences with other statistical programs can be observed.

Figure IX.12. Beanplot of the median and distribution of nitrite concentration (μML) in two lakes of the Amazon.



Value

A TXT file is obtained with the results of the test and a boxplot or beanplot can be displayed.

References

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28: 1-9.

Kampstra, P (2013) Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot). R package version 1.1. Available at: http://CRAN.R-project.org/package=beanplot.

Pérez, C (2004) *Técnicas de análisis multivariante de datos*. Aplicaciones con SPSS. Pearson Educación. Madrid.

Sokal, R.R. & Rohlf, F.J. (1981) *Biometry*. WH Freeman and Company, New York.

Zar, J.H. (1999) *Biostatistical Analysis*. Prentice Hall, New Jersey.

Examples

```
## Not run:

data(ZIX7)

IX9(data=ZIX7, variables="Nitrite", factor="Lake", pop1="Correo", pop2="Tarapoto",
graph="Beanplot")


## End(Not run)
```

PAR *PAR FUNCTION*

Description

This function allows to modify graphs.

Usage

```
PAR(adj=0.5, ann=TRUE, bg=NULL, bty="o", cex.axis=1, cex.lab=1, cex.main=1,
cex.sub=1, col.axis="black", col.lab="black", col.main="black",
col.sub="black", family="Arial", fg="black", font.axis=1, font.lab=1,
font.main=2, font.sub=1, lab=c(5,5,7), las=0, lend=0, lty= 1, lwd=1, mai=NULL,
mar=NULL, mex=NULL, mgp=c(3,3,0), oma=NULL, omd=NULL, omi=NULL, pch=NULL,
ps=NULL, pty="m", tck=NA, tcl=NA, xaxp=NULL, xaxs="r", xaxt="s", yaxp=NULL,
yaxs="r", yaxt="s")
```

Arguments

adj	It specifies the alignment of the text in graphics. It can take any value between 0 (text to the left) and 1 (text to the right).
ann	If you select FALSE, the legends of the axes and the titles of the graphics are removed when used with the function plot.default.
bg	Background color of the graph.

`bty` It defines the type of the graphics frame:
"o" (full frame)
"c" (the right edge is not shown)
"u" (the upper edge is not shown)
"l" (the top and right edges are not shown)
"7" (left and bottom edges are not shown)
"]" (the left edge is not shown).

`cex.axis` Size of the axis labels.

`cex.lab` Size of the text of the legends.

`cex.main` Size of the graph title text.

`cex.sub` Size of the text in the subtitle.

`col.axis` It defines the color of the axis labels.

`col.lab` It defines the color of the text of legends.

`col.main` It defines the color of the text of the title of the graph.

`col.sub` It defines the color of the text in the subtitle.

`family` It specifies the font of the text.

`fg` It defines the frame color and lines of the graph.

`font.axis` A numeric value that defines the font of the axis labels. The value 1 is a normal type, 2 is written in bold, 3 is written in italics and 4 is written in italics and bold.

`font.lab` A numeric value that defines the font of the legends. The value 1 is a normal type, 2 is written in bold, 3 is written in italics and 4 is written in italics and bold.

`font.main` A numeric value that defines the font of the title of the graph. The value 1 is a normal type, 2 is written in bold, 3 is written in italics and 4 is written in italics and bold.

`font.sub` A numeric value that defines the font of the subtitle of the graph. The value 1 is a normal type, 2 is written in bold, 3 is written in italics and 4 is written in italics and bold.

`lab` A numeric value with the format c(x,y,len) that defines the number of axis divisions. The value of «len» must be specified but R does not use it.

`las` It defines the position of the axis labels: 0 is parallel to the shaft, 1 is horizontal, 2 is perpendicular, and 3 is vertical.

`lend` It defines the style of the end of the line and it can be specified with a number or letters:
0 or "round" is a rounded end.
1 or "butt" is a thick end.
2 or "square" is a square end.

`lty` It defines the type of line:
0 No line.
1 Solid line.

2 Dashed line.

3 Dotted line.

4 Line of dots and dashes.

5 Dash line.

6 Double stripe.

`lwd` It defines the line width.

`mai` A numeric vector with the format c(down, left, up, right) that defines the margins of the figure in inches.

`mar` A numeric vector with the format c(down, left, up, right) that defines the lines of the margins of the figure.

`mex` Expansion factor of the size of characters on the margins.

`mgp` A numeric vector with the format c(t, e, l) that specifies the position of the legend of the axes (t), the text of the divisions of the axes (e) and the line of axis (l). The default value is (3,1,0).

`oma` A numeric vector with the format c(down, left, up, right) that defines the margins of the text in the panel, in number of lines.

`omd` A numeric vector with the format c(x1,x2,y1,y2) that defines the internal margins where the graph will be located on a scale from 0 to 1.

`omi` A numeric vector with the format c(down, left, up, right) that defines the margins of the panel in inches.

`pch` Numeric value between 1 and 25 that defines the type of symbol (the symbol that corresponds to each number is shown in the table below). It is also possible to put "" between any character, as shown in the table with a, *, !, ¡ y $.

`ps` Text point size (unit of measurement of the characters) in 1/72 inches.

`pty` Character that indicates whether the graph is square or stretches laterally when the window is enlarged. With "s" produces a square figure and with "m" the graph is extended to the maximum extent possible.

`tck` String value that defines the size of the marks on the axes, in proportion to the area of the region of the figure. A value of 1 puts inner lines in the graph. With tck = NA it leaves the mark by default.

`tcl` String value that defines the size of the marks on the axes. It differs from *tc* in which the measures are real and not in proportion. With *tcl=NA* leaves the mark by default.

`xaxp` Numeric vector with the format c(x1, x2, n), where x1 and x2 are the ends of the marks of x-axis, and n is the number of intervals or marks on the shaft.

`xaxs` "r" (regular) extends 4% more than the limits given in «xlim», and calculates the number of intervals that best fit those limits, and "i" (internal) seeks most appropriate number of intervals to the original data.

`xaxt` Logical value that specifies whether to represent the intervals of the x-axis, with the options "s" or "n".

`yaxp` Vector numeric format c (x1, x2, n), where x1 and x2 are the extremes axis marks Y, and n the number of intervals or marks on the shaft.

yaxs
: r" (regular) extends 4% more than the limits given in «ylim», and calculates the number of intervals that best fit those limits, and "i" (internal) seeks most appropriate number of intervals to the original data.

yaxt
: Logical value that specifies whether to represent the intervals of the y-axis, with the options "s" or "n".

VI1 *CONTRAST OF GOODNESS OF FIT-POISSON*

Description

It is determined if a sample is adjusted to a Poisson distribution.

Usage

```
VI1(data, variables, group=NULL, distribution="Normal", combine=TRUE,
graph=TRUE, ni=12, xlab="", col.bar="cyan", col.line="red", lty=2,
lwd=2, legend.pos="topleft", file="Output.csv", na="NA", dec=",",
row.names=FALSE)
```

Arguments

data
: Data file.

variables
: Variable or variables for which the contrast of goodness of fit can be performed.

group
: Variables that gather data for calculations. In case of selecting NULL, there would be no grouping and it would be calculated considering all the data of the selected variables.

distribution
: Type of theoretical distribution which contrasts the adjustment of the sample: "Binomial", "Poisson" or "Normal".

combine
: It is advisable that if the number of expected cases within a class is less than 5, several categories can be combined in one until all have an expected frequency greater than or equal to 5. For this reason, it is convenient to leave the default value TRUE.

graph
: GRAPHIC. Logical value which defines whether it represents the graph with the expected and actual distributions.

ni
: GRAPHIC. Number of bars (intervals) in the graph of the Normal distribution.

xlab
: GRAPHIC. Legend of the X-axis.

col.bar
: GRAPHIC. Color of the bars.

col.line
: GRAPHIC. Color of the line of the theoretical distribution.

lty
: GRAPHIC. Type of the theoretical distribution line.

lwd
: GRAPHIC. The line width of the theoretical distribution.

legend.pos
: GRAPHIC. Position of the legend on the chart: "topleft", "topright", "bottomleft" o "bottomright"

`file` CSV FILE. Output file name.

`na` CSV FILE. Text used in cells without data.

`dec` CSV FILE. Defines whether the comma "," or dot "." is used as decimal separator.

`row.names` CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

VI. CONTRAST OF GOODNESS OF FIT

To determine if the variables are adjusted to a Normal distribution, Poisson, Binomial, etc., it is necessary to use statistical contrasts, which are called goodness-of-fit tests. There are different goodness-of-fit tests that are used depending on the type of data and the theoretical distribution expected. A classification of the settings used is:

1. Categorized sample (data are codes assigned to the values of a qualitative variable or class in which a quantitative variable values are grouped): Chi-square

2. Samples not categorized (quantitative variables, continuous or discrete, not grouped into intervals or classes).

a) For all distributions: Kolmogorov-Smirnov test (test K-S).

b) Normal Distribution: Test of normality Shapiro-Wilk.

VI.1. CHI-SQUARE

This applies to continuous distributions (with data previously grouped in classes) as to discrete distributions or qualitative variables. It is based on quantifying the differences between the observed frequencies in each class and the expected ones, based on the null hypothesis that the data follow an *F(x)* distribution (which can be Normal, Poisson, etc.). For its application in the *n* existing classes, the number of observed cases (O_i) to be contrasted and, through, the theoretical function, the number of expected cases (E_i) is calculated. From these frequencies, the statistical value of χ^2, is calculated using the following formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

It is recommended that if the number of expected cases within a class is less than 5, various categories are combined in one, until all have an expected frequency greater than or equal to 5. It should not be used when there are few data. Then the degrees of freedom (υ) of the sample should be calculated. If the expected values can be calculated prior to sampling, the number of υ is *n* - 1. However, if to calculate the expected values is necessary to estimate some statistics obtained in the sampling parameters (such as, μ or σ for a Normal distribution), the number of υ is *n* - *r* -1 where *r* is the number of statistics necessary to calculate the expected value.

Once determined χ^2 and υ, it is necessary to search in the table χ^2 (Table 3 Appendix I Guisande et al., 2011) the critical value for these υ and to the chosen level of significance (usually α = 0.05). If the χ^2 critical is greater than χ^2 calculated, the null hypothesis that the observed data are consistent with the expected distribution is accepted, while if χ^2 critical is less than χ^2 calculated, the hypothesis is rejected. Most of the statistical programs directly calculate the value *p* of the contrast; if it is less than the chosen level of significance (usually 0.05 or 0.01), the hypothesis is rejected.

VI.2. KOLMOGOROV-SMIRNOV TEST

This is the correct test to compare the normality of a sample if the number of data is large (*n* > 30), although it can be used for both large and small samples. It can also be used to compare other distributions such as the Binomial or Poisson. It is a very conservative test that is applied to continuous variables. It is based on the determination of the maximum (*D*) difference between the observed cumulative frequencies (AO_i) and the expected cumulative frequencies (AE_i), based on the null hypothesis that the data follow a certain distribution. The formula of the test is $D = max|AO_i - AE_i|$. This test was recalculated to a Normal distribution studying the expected frequencies based on the mean and variance of the sample (Lilliefors, 1967) and is known as test *K-S-L*. However, its application is limited when there are few data (100 observations are necessary to distinguish between a Normal with μ = 0 and σ^2 = 1 and a uniform distribution between $-\sqrt{3}$ and $\sqrt{3}$.

Once the statistical *D* is calculated, it is compared with a *D* critical value for the chosen level of significance that appears in Table 5 of Appendix I (Guisande et al. , 2011) and Table 6 of Appendix I (Guisande et al. , 2011, for the *K-S-L* test). The null hypothesis is accepted when *D*observed is lower than the value of *D* tabulated or when the value *p* of the contrast is greater that the level of significance (for example greater than 0.05).

VI.3. SHAPIRO-WILK TEST

It is the recommended test to verify the normality of a sample, especially when working with a small number of data (*n* < 30). It is only used to verify the Normal distribution (Shapiro & Wilk, 1965). It is based on measuring the adjustment of data to a straight Normal probability (Figure VI.1). If the setting were perfect, the points would form a straight line of 45º (observed frequency equal to expected frequency). The statistic test is expressed by using the following equation:

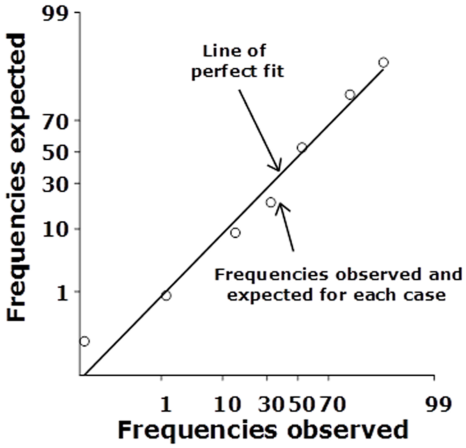
$$W = \frac{1}{\sum_{j=1}^{n}(x_j - \mu)^2}\left[\sum_{j=1}^{h} a_{j,n}(x_{(n-j+1)} - x_j)\right]^2$$

where *n* is the number of data, x_j is the sample data in ascending order which ranks *j*, μ is the mean, *h* is *n*/2 if *n* is even or (*n*-1)/2 if *n* is odd and $a_{j,n}$ is a tabulated value.

Once the statistical *W* is calculated, this is contrasted with a critical value *W* for the chosen level of significance.

As this statistic measures the adjustment to a straight line and not the distance to the normal distribution (as it was in previous cases),and can be interpreted in an approximate way as a correlation coefficient between the observed and expected values (a value close to 1 indicates a good fit, and next to 0 a bad fit), the null hypothesis is accepted when the value *W* is higher than the tabulated value of contrast (very high adjusted value).

Figure VI.1. Representation of expected and observed probabilities of a sample along with the line of best fit to a Normal distribution.

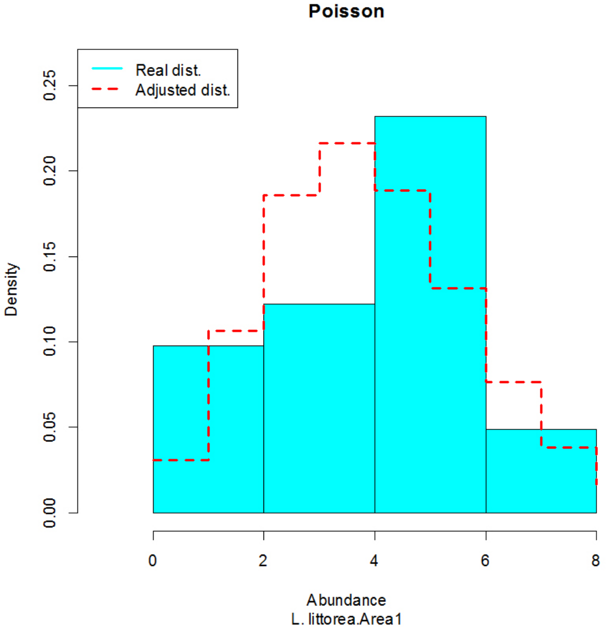


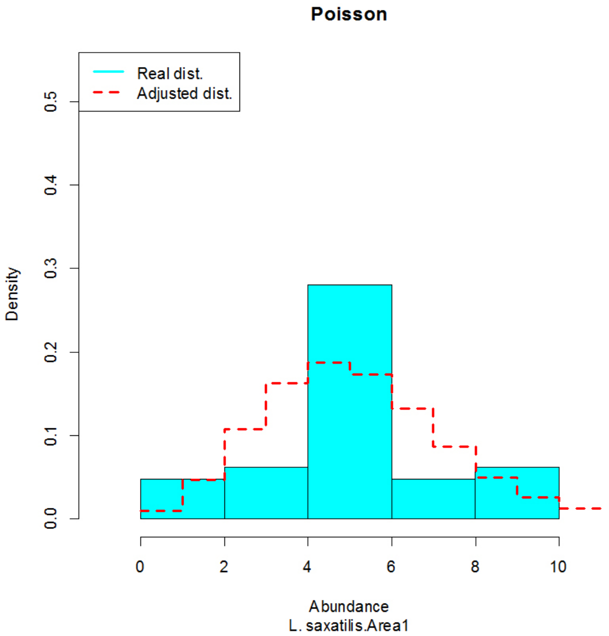
EXAMPLE

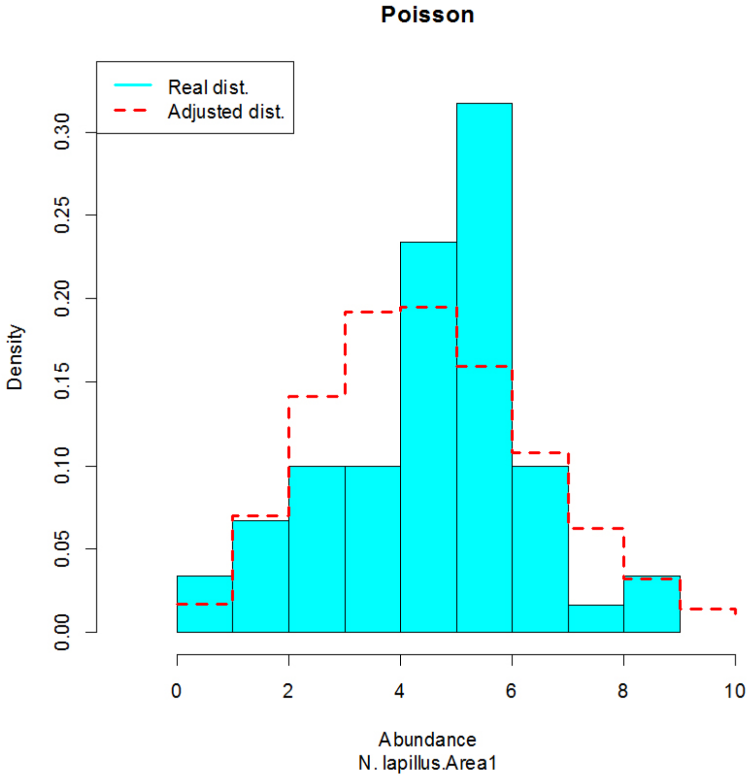
Data from the abundance of two herbivorous species of sea snails (*Littorina littorea* and *Littorina saxatilis*) and a carnivorous species (*Nucella lapillus*), in two areas of sampling. It is necessary to know if the distribution is random in different areas, i.e., if this follows a Poisson distribution.

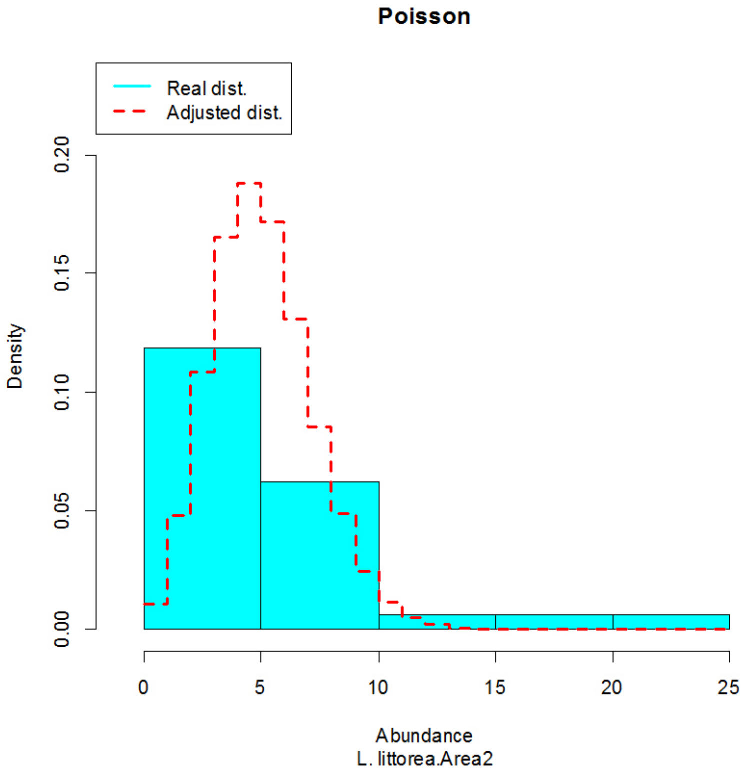
To see all the charts, which are shown below (Figure VI.2), simply drag each of the windows, and thus see the graph below that.

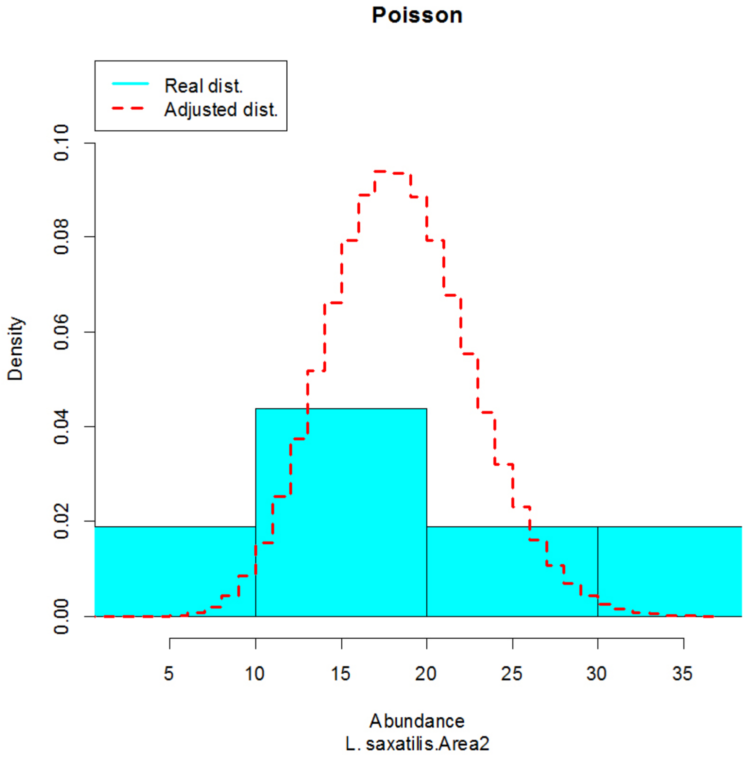
Figure VI.2. Frequency histograms for each of the species in the two areas.

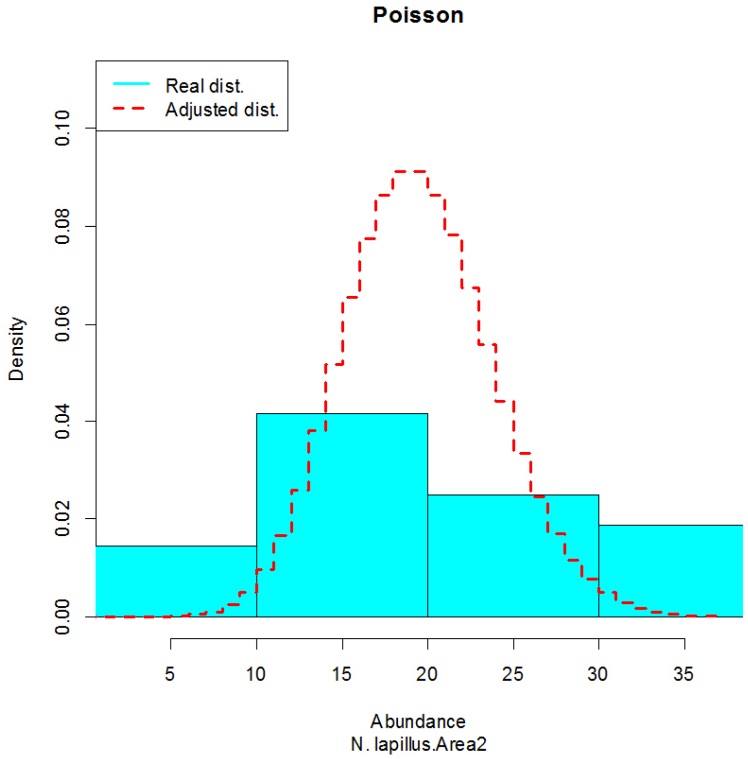



Poisson
Real dist.
Adjusted dist.
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
Abundance
L. saxatilis.Area1

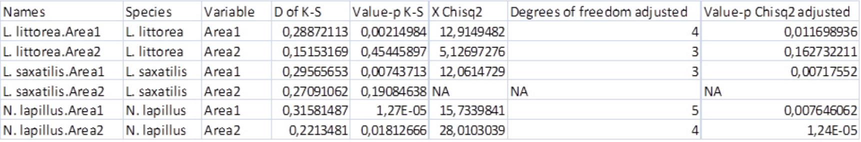

Poisson
Real dist.
Adjusted dist.
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10
Abundance
N. lapillus.Area1


Poisson
Real dist.
Adjusted dist.
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
Abundance
L. littorea.Area2


Poisson
Real dist.
Adjusted dist.
Density
0.00
0.02
0.04
0.06
0.08
0.10
5
10
15
20
25
30
35
Abundance
L. saxatilis.Area2



The results are shown below:



Names	Species	Variable	D of K-S	Value-p K-S	X Chisq2	Degrees of freedom adjusted	Value-p Chisq2 adjusted
L. littorea.Area1	L. littorea	Area1	0,28872113	0,00214984	12,9149482	4	0,011698936
L. littorea.Area2	L. littorea	Area2	0,15153169	0,45445897	5,12697276	3	0,162732211
L. saxatilis.Area1	L. saxatilis	Area1	0,29565653	0,00743713	12,0614729	3	0,00717552
L. saxatilis.Area2	L. saxatilis	Area2	0,27091062	0,19084638	NA	NA	NA
N. lapillus.Area1	N. lapillus	Area1	0,31581487	1,27E-05	15,7339841	5	0,007646062
N. lapillus.Area2	N. lapillus	Area2	0,2213481	0,01812666	28,0103039	4	1,24E-05

In the case of the test χ^2, the test set comprises the categories to achieve expected frequencies no less than five. When the results of the unadjusted and adjusted tests are different, this means that it has been necessary to group them and, therefore, the unadjusted test is invalid. The results show that in all cases the probability values for test χ^2 of the adjusted test are different from the unadjusted and therefore, in all cases the adjusted test must be considered as valid.

For *Littorina littorea* in zone 1, the value *p* of the contrast χ^2 is less than 0.05 in the adjusted test (p = 0.011), therefore, the hypothesis that the distribution of this species is random in zone 1 is rejected, i.e. it does not follow a Poisson distribution. The Kolmogorov-Smirnov test (*K-S*) corroborates the before mentioned since p = 0.002 (is less than 0.05). On the contrary, for the same species in zone 2, the value *p* of the contrast χ^2 is greater than 0.05 in the adjusted test (p = 0.162), indicating that there can be a Poisson distribution in this area. The test *K-S* confirms what is already mentioned before that p = 0.454 (is greater than 0.05).

In the case of *Littorina saxatilis*, the value *p* of the contrast χ^2 is less than 0.05 in zone 1 (p = 0.007 adjusted), indicating that it does not follow a Poisson distribution, and can not be estimated in zone 2.

With the test of *K-S* the probability is also less than 0.05 in zone 1 (p = 0.007), indicating that the distribution is not Poisson. The test *K-S* is more appropriate than the test χ^2 for quantitative variables, so that we would give validity to the test *K-S*. In any case, if you decide to apply multiple valid evidence for the same purpose, as in this example where we use the χ^2 and *K-S* tests, and some of them reject the null hypothesis and some not, the null hypothesis must be rejected.

For *Nucella lapillus*, in the two areas, both the χ^2 of the adjusted tests as the *K-S* test, the p-value < 0.05 , rejecting the hypothesis of random Poisson distribution.

Value

A CSV file is obtained with the results of the contrasts and the graph with the expected and actual distributions if the option to show was selected.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Lilliefors, H. (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62: 399-402.

Shapiro, S.S. & Wilk, M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, 52: 591-611.

Examples

```
## Not run:

data(ZVI1)

#Poisson distribution
VI1(data=ZVI1, variables=c("Area1","Area2"), group =c("Species"),
distribution="Poisson", xlab="Abundance")


## End(Not run)
```

VI2 *CONTRAST OF GOODNESS OF FIT-BINOMIAL*

Description

It determines whether a sample is adjusted to a Binomial distribution.

Usage

```
VI2(data, variables, group=NULL, distribution="Normal", combine=TRUE,
graph=TRUE, ni=12, xlab="", col.bar="cyan", col.line="red", lty=2, lwd=2,
legend.pos="topleft", file="Output.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data` Data file.

`variables` Variable or variables for which the contrast of goodness of fit can be performed.

`group` Variables that gather data for calculations. In case of selecting NULL, there would be no grouping and it would be calculated considering all the data of the selected variables.

`distribution` Type of theoretical distribution which contrasts the adjustment of the sample: "Binomial", "Poisson" or "Normal".

`combine` It is advisable that if the number of expected cases within a class is less than 5, several kinds can be combined in one until all have an expected frequency greater than or equal to 5. For this reason, it is convenient to leave the default value TRUE.

`graph` GRAPHIC. Logical value which defines whether it represents the graph with the expected and actual distributions.

`ni` GRAPHIC. Number of bars (intervals) in the graph of the Normal distribution.

`xlab` GRAPHIC. Legend of the X-axis.

`col.bar` GRAPHIC. Color of the bars.

`col.line` GRAPHIC. Color of the line of the theoretical distribution.

`lty` GRAPHIC. Type of the theoretical distribution line.

`lwd` GRAPHIC. Line-width of the theoretical distribution.

`legend.pos` GRAPHIC. Position of the legend on the chart: "topleft", "topright", "bottomleft" o "bottomright"

`file` CSV FILE. Output file name.

`na` CSV FILE. Text used in the cells without data.

`dec` CSV FILE. It defines whether the comma "," or dot "." is used as decimal separator.

`row.names` CSV FILE. Logical value that defines if identifiers are placed in rows or a vector with a text for each of the rows.

Details

VI. CONTRAST OF GOODNESS OF FIT

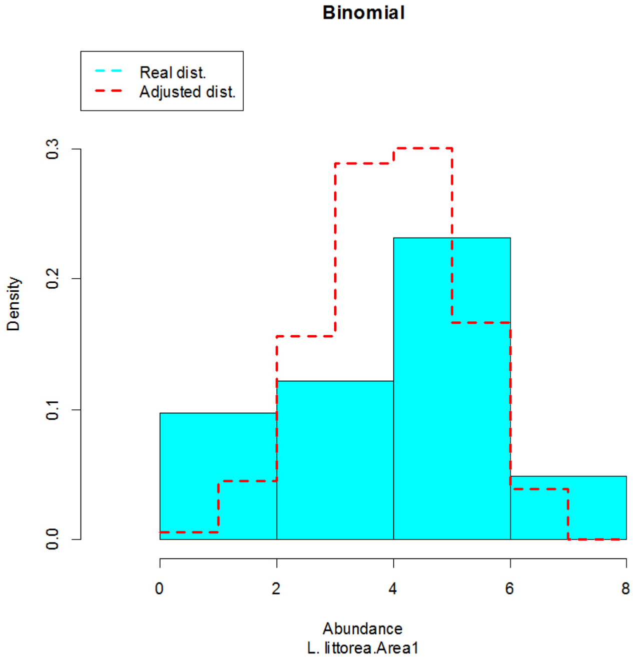
See section *details* of the function VI1.

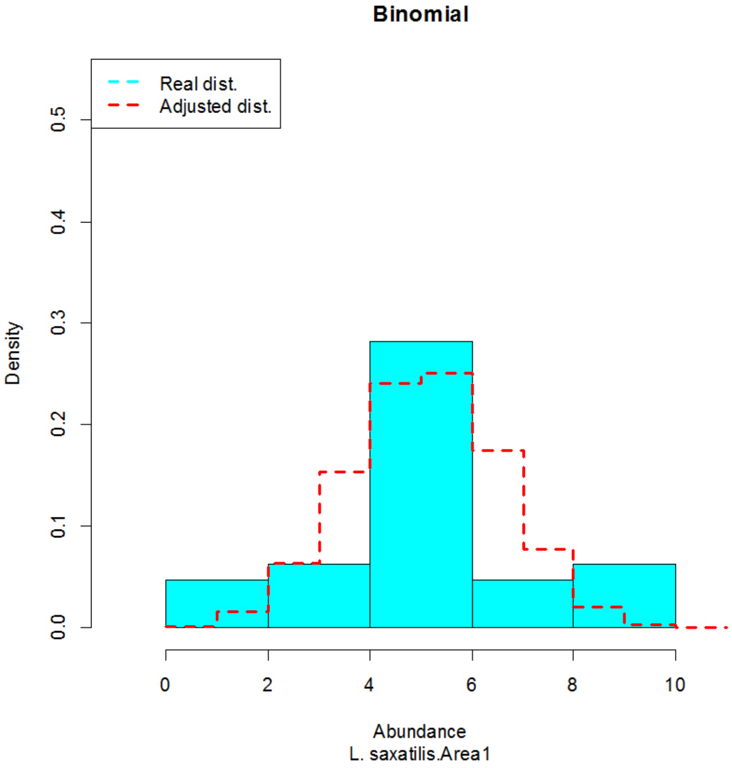
EXAMPLE

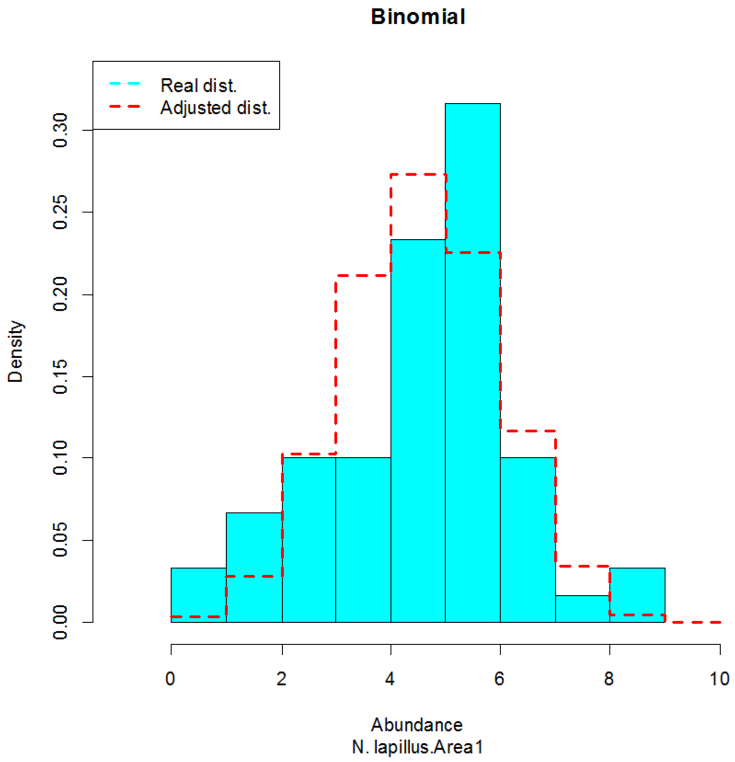
Data from the abundance of two herbivorous species of sea snails (*Littorina littorea* and *Littorina saxatilis*) and a carnivorous species (*Nucella lapillus*), in two sampling areas. It is necessary to know if the distribution is Binomial in both areas and for each species.

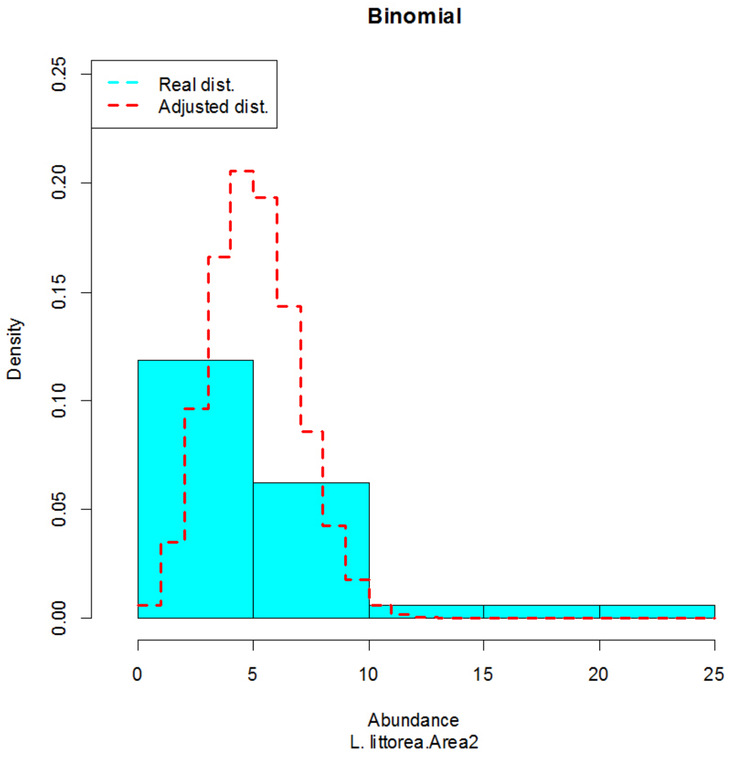
To view all the charts, which are shown below (Figure VI.3), simply drag each of the windows, and thus, the chart below can be seen.

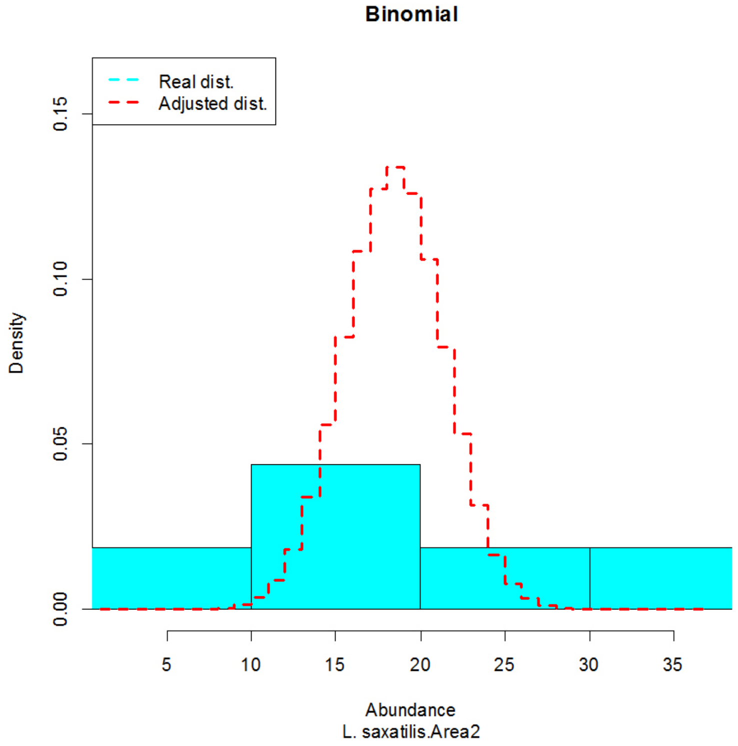
Figure VI.3. Frequency histograms for each of the species
in the two areas.

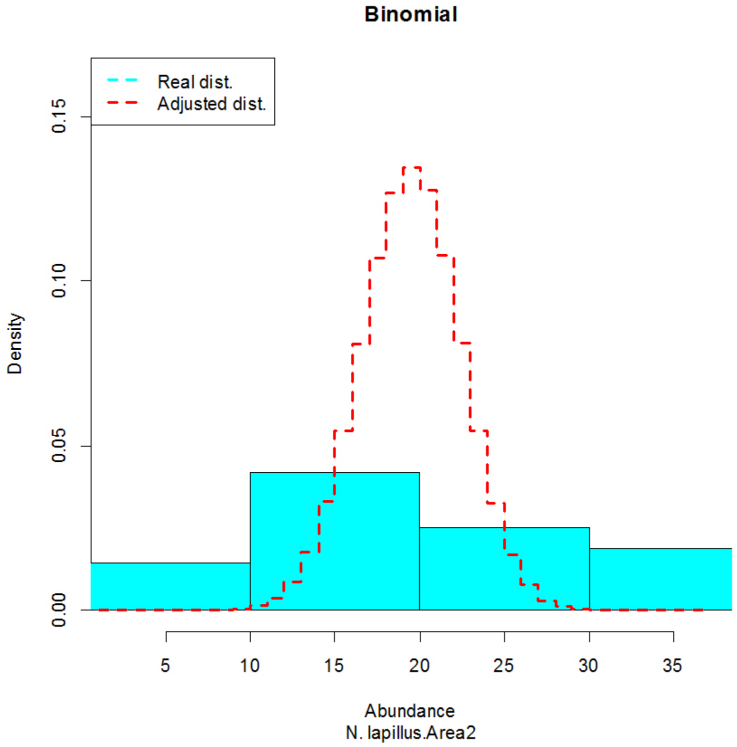

Binomial
Real dist.
Adjusted dist.
Density
0.0
0.1
0.2
0.3
0
2
4
6
8
Abundance
L. littorea.Area1


Binomial
Real dist.
Adjusted dist.
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
Abundance
L. saxatilis.Area1

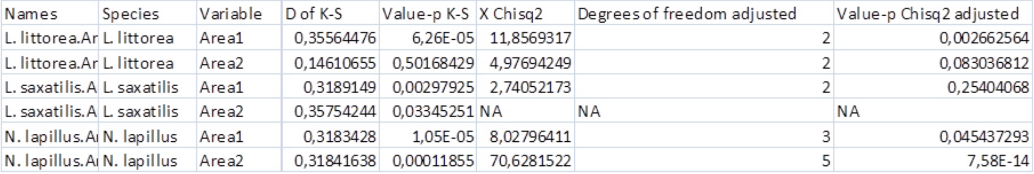

Binomial
Real dist.
Adjusted dist.
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10
Abundance
N. lapillus.Area1


Binomial
Real dist.
Adjusted dist.
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20
25
Abundance
L. littorea.Area2


Binomial
Real dist.
Adjusted dist.
Density
0.00
0.05
0.10
0.15
5
10
15
20
25
30
35
Abundance
L. saxatilis.Area2


Binomial
Real dist.
Adjusted dist.
Density
0.00
0.05
0.10
0.15
5
10
15
20
25
30
35
Abundance
N. lapillus.Area2

The results are shown below:



Names	Species	Variable	D of K-S	Value-p K-S	X Chisq2	Degrees of freedom adjusted	Value-p Chisq2 adjusted
L. littorea.Ar	L. littorea	Area1	0,35564476	6,26E-05	11,8569317	2	0,002662564
L. littorea.Ar	L. littorea	Area2	0,14610655	0,50168429	4,97694249	2	0,083036812
L. saxatilis.A	L. saxatilis	Area1	0,3189149	0,00297925	2,74052173	2	0,25404068
L. saxatilis.A	L. saxatilis	Area2	0,35754244	0,03345251	NA	NA	NA
N. lapillus.A	N. lapillus	Area1	0,3183428	1,05E-05	8,02796411	3	0,045437293
N. lapillus.A	N. lapillus	Area2	0,31841638	0,00011855	70,6281522	5	7,58E-14

In the case of the test χ^2, the test set comprises the categories to achieve expected frequencies no less than five. When the results of the unadjusted and adjusted tests are different, this means that it has been necessary to group them and, therefore, the unadjusted test is invalid. The results show that in all cases the probability values for test χ^2 of the adjusted test are different from the unadjusted and therefore, in all cases the adjusted test must be considered as valid.

When $p < 0.05$ the null hypothesis is rejected and, therefore, the distribution of the species does not follow a Binomial one. On the contrary, in the case of *L. littorea* in zone 2 ($p = 0.083$) and *L. saxatilis* in zone 1 ($p = 0.254$) species show a Binomial distribution.

Value

A CSV file is obtained with the results of the contrasts and the graph with the expected and actual distributions if the option to show was selected.

Examples

```
## Not run:

data(ZVI1)

#Binomial distribution
VI2(data=ZVI1, variables=c("Area1","Area2"), group=c("Species"),
distribution="Binomial", xlab="Abundance")

## End(Not run)
```

VI3 *CONTRAST OF GOODNESS OF FIT-NORMAL*

Description

It determines if a sample is adjusted to a Normal distribution.

Usage

```
VI3(data, variables, group=NULL, distribution="Normal", combine=TRUE,
graph=TRUE, ni=12, xlab="", col.bar="cyan", col.line="red", lty=2, lwd=2,
legend.pos="topleft", file="Output.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data` Data file.

`variables` Variable or variables for which the contrast of goodness of fit can be performed.

`group` Variables that gather data for calculations. In case of selecting NULL, there would be no grouping and it would be calculated considering all the data of the selected variables.

`distribution` Type of theoretical distribution which contrasts the adjustment of the sample: "Binomial", "Poisson" or "Normal".

`combine` It is advisable that if the number of expected cases within a class is less than 5, several kinds can be combined in one until all have an expected frequency greater than or equal to 5. For this reason, it is convenient to leave the default value TRUE.

`graph` GRAPHIC. Logical value which defines whether it represents the graph with the expected and actual distributions.

`ni` GRAPHIC. Number of bars (intervals) in the graph of the Normal distribution.

`xlab` GRAPHIC. Legend of the X-axis.

`col.bar` GRAPHIC. Color of the bars.

`col.line` GRAPHIC. Color of the line of the theoretical distribution.

`lty` GRAPHIC. Type of the theoretical distribution.

`lwd` GRAPHIC. Line-width of the theoretical distribution.

`legend.pos` GRAPHIC. Position of the legend on the chart: "topleft", "topright", "bottomleft" o "bottomright"

`file` CSV FILE. Output file name.

`na` CSV FILE. Text used in the cells without data.

`dec` CSV FILE. It defines whether the comma "," or dot "." is used as decimal separator.

`row.names` CSV FILE. Logical value that defines if identifiers are placed in rows or a vector with a text for each of the rows.

Details

VI. CONTRAST OF GOODNESS OF FIT

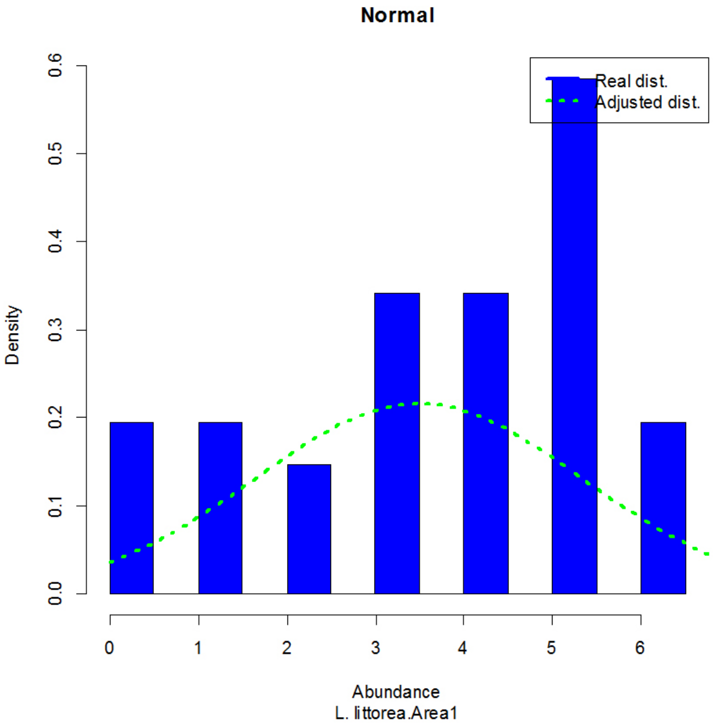
See section *details* of the function VI1.

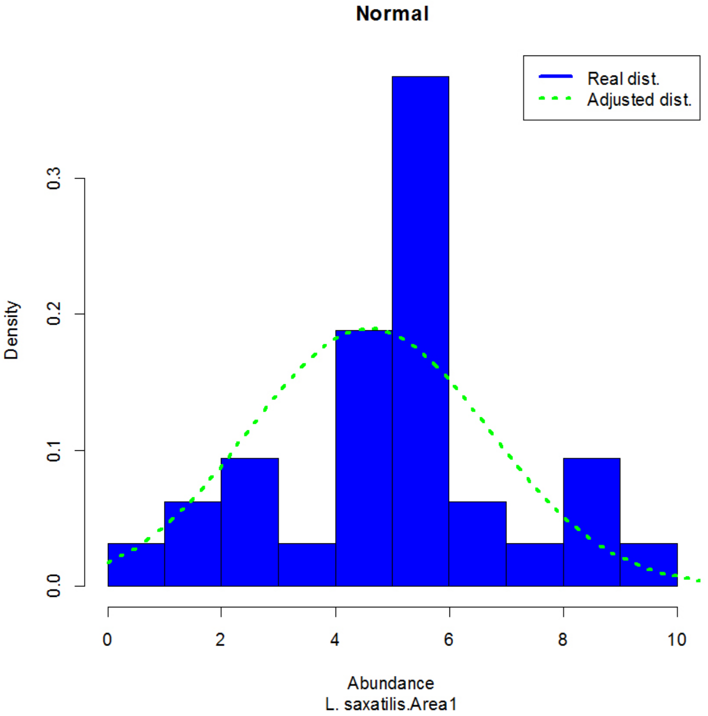
EXAMPLE

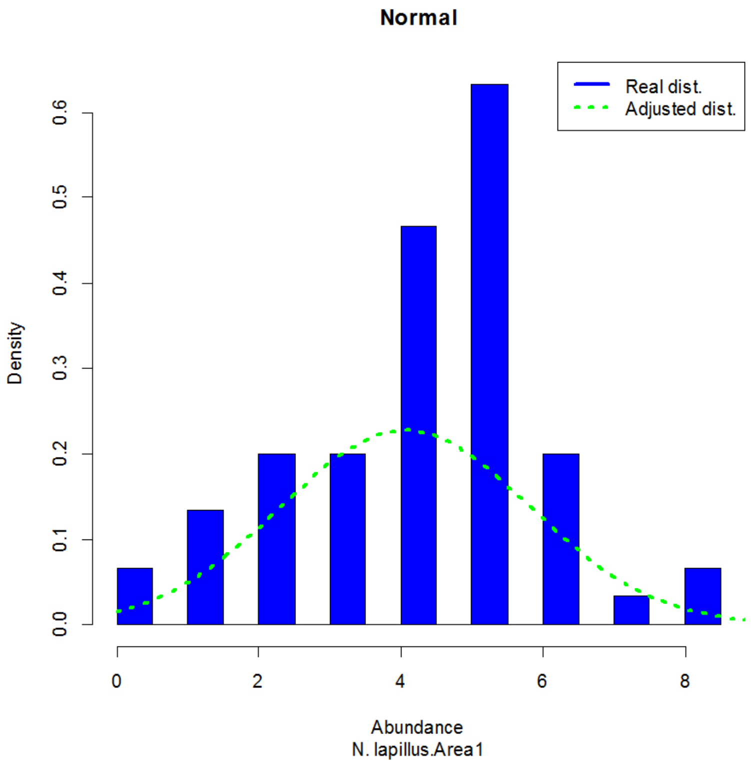
Data from the abundance of two herbivorous species of sea snails (*Littorina littorea* and *Littorina saxatilis*) and a carnivorous species (*Nucella lapillus*), in two sampling areas. It is necessary to know if the distribution is Normal in both areas and for each species.

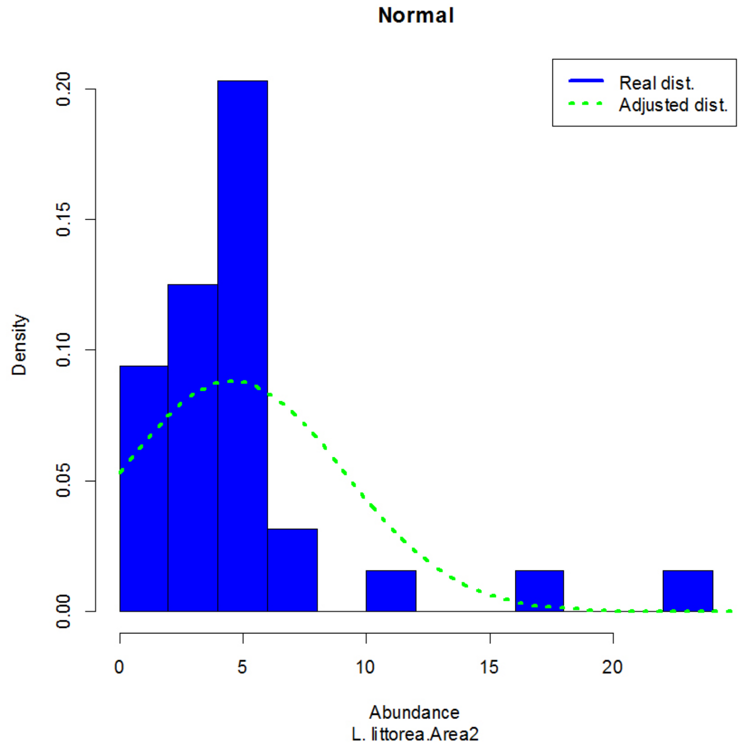
To view all graphs, which are shown below (Figure VI.4), simply drag each of the windows, and thus, the chart below can be seen.

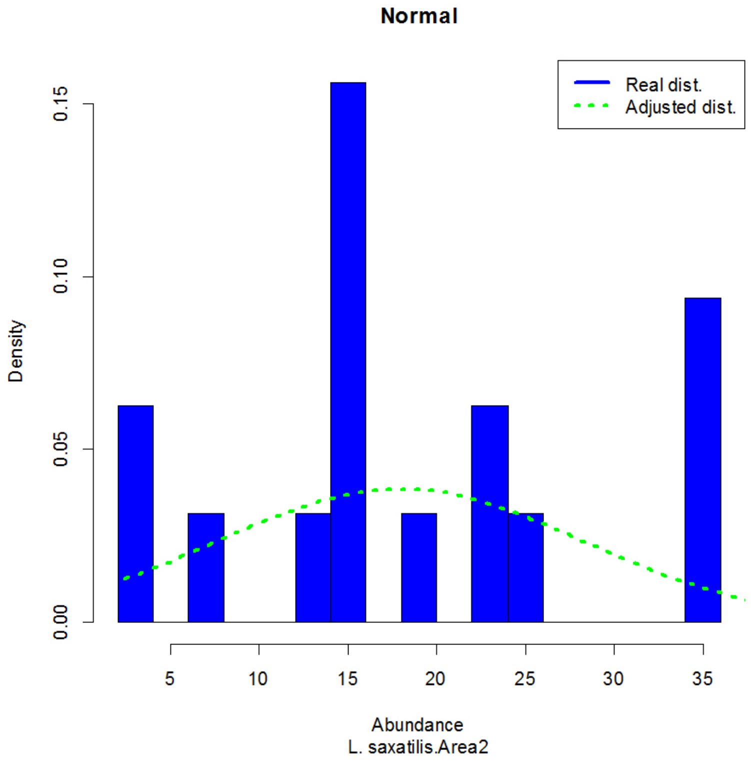
Figure VI.4. Frequency histograms for each of the species
in the two areas.

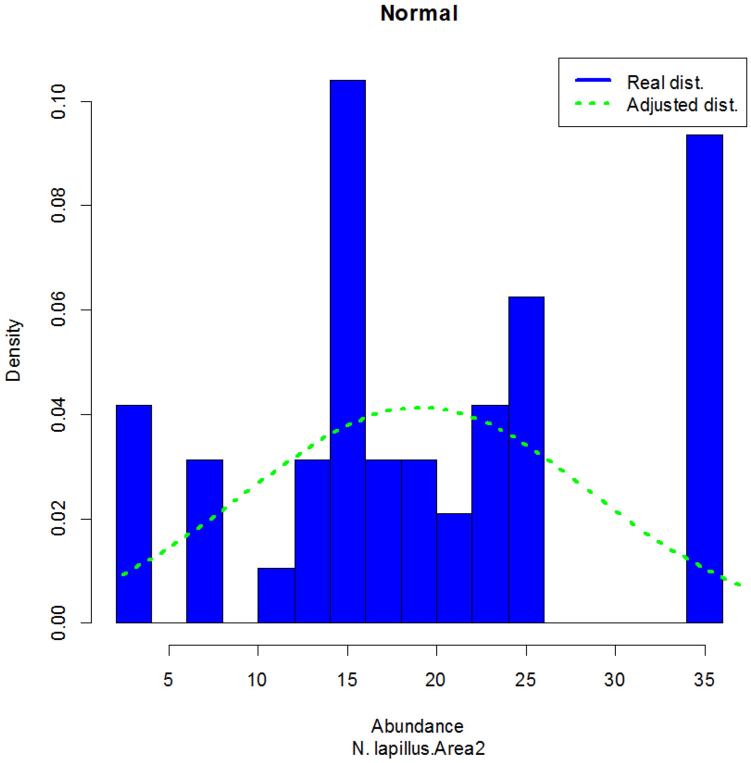

Normal
Real dist.
Adjusted dist.
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
5
6
Abundance
L. littorea.Area1


Normal
Real dist.
Adjusted dist.
Density
0.0
0.1
0.2
0.3
0
2
4
6
8
10
Abundance
L. saxatilis.Area1

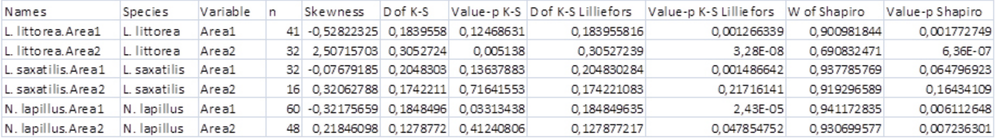

Normal
Real dist.
Adjusted dist.
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
Abundance
N. lapillus.Area1


Normal
Real dist.
Adjusted dist.
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
Abundance
L. littorea.Area2


Normal
Real dist.
Adjusted dist.
Density
0.00
0.05
0.10
0.15
5
10
15
20
25
30
35
Abundance
L. saxatilis.Area2


Normal
Real dist.
Adjusted dist.
Density
0.00
0.02
0.04
0.06
0.08
0.10
5
10
15
20
25
30
35
Abundance
N. lapillus.Area2

The results are shown below:



Names	Species	Variable	n	Skewness	D of K-S	Value-p K-S	D of K-S Lilliefors	Value-p K-S Lilliefors	W of Shapiro	Value-p Shapiro
L. littorea.Area1	L. littorea	Area1	41	-0,52822325	0,1839558	0,12468631	0,183955816	0,001266339	0,900981844	0,001772749
L. littorea.Area2	L. littorea	Area2	32	2,50715703	0,3052724	0,005138	0,30527239	3,28E-08	0,690832471	6,36E-07
L. saxatilis.Area1	L. saxatilis	Area1	32	-0,07679185	0,2048303	0,13637883	0,204830284	0,001486642	0,937785769	0,064796923
L. saxatilis.Area2	L. saxatilis	Area2	16	0,32062788	0,1742211	0,71641553	0,174221083	0,21716141	0,919296589	0,16434109
N. lapillus.Area1	N. lapillus	Area1	60	-0,32175659	0,1848496	0,03313438	0,184849635	2,43E-05	0,941172835	0,006112648
N. lapillus.Area2	N. lapillus	Area2	48	0,21846098	0,1278772	0,41240806	0,127877217	0,047854752	0,930699577	0,007236301

The results also show the asymmetry of each of the data sets since under the assumption that the population does not have a Normal distribution, it is possible to transform data and the asymmetry indicates the type of transformation to be performed (see section *details* of the function IX1).

There is only a discrepancy when comparing the test *K-S* with the Shapiro-Wilk test. The species *L. saxatilis* in zone 1 does not have a Normal distribution as the test *K-S* (p = 0.001), while the Shapiro-Wilk test shows a normal distribution (p = 0.065).

As mentioned before, if there is a discrepancy between two tests, a conservative decision must be made, that is to say, rejecting the null hypothesis. In this example the Normal distribution is rejected.

Value

A CSV file is obtained with the results of the contrasts and the graph with the expected and actual distributions if the option to show it was selected.

Examples

```
## Not run:

data(ZVI1)

#Normal distribution
VI3(data=ZVI1, variables=c("Area1","Area2"), group=c("Species"),
xlab="Abundance", col.bar="blue", col.line="green", lty=3, lwd=3,
legend.pos="topright")


## End(Not run)
```

VII1 *CONTRASTS OF HOMOGENEITY IN QUALITATIVE VARIABLES - INDEPENDENT SAMPLES AND POLYTOMOUS VARIABLES*

Description

It determines if multiple samples taken from the same population or from different populations, differ in a certain qualitative variable.

Usage

```
VII1(data, variables, group=NULL, test=c(1,2,3), graph=TRUE, compress=TRUE,
xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, ”lines”), main=NULL,
sub=NULL, residuals_type=”Pearson”, shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col=”black”), grid.edit=FALSE, text.grid=”rect:”, gp.grid = gpar(fill=”red”),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, ”lines”),
legend_width=unit(5, ”lines”), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily=”Arial”, x= unit(1, ”lines”), y=unit(0.1,”npc”), height=unit(0.8, ”npc”),
width=unit(0.7, ”lines”), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file=”Output.txt”)
```

Arguments

`data` Data file.

`variables` Variable or variables to which the contrast of homogeneity is to be performed.

`group` Variables together with the data are grouped for calculations. In case of selecting NULL, no grouping would be done, and this would be calculated considering all data of the selected variables.

`test` TYPE OF CONTRAST OF HOMOGENEITY.

1. Chi-square test.
2. G-test without the Williams' correction.
3. G-test with the Williams' correction.
4. Chi-square test with Yates correction (Only for 2x2 tables).
5. Fisher's exact test (Only for 2x2 tables), bilateral.
6. Fisher's exact test (Only for 2x2 tables), unilateral: less than.
7. Fisher's exact test (Only for 2x2 tables), unilateral: greater than.
8. Friedman ANOVA test.
9. McNemar's test without the continuity correction.
10. McNemar's test with the continuity correction (Only for 2x2 tables).
11. Cochran test.

`graph` GRAPH. If TRUE, the graph of the contingency table is shown. This chart does not allow accents and the categories are sorted by alphabetical order. The example of the function VIII1 shows how to sort the categories.

`compress` GRAPH. Logical value, if FALSE the space between rows and columns is chosen so that the total of heights and widths of the rows and columns are equal. If TRUE, the space between rows and columns is fixed and, therefore, the graph is more compressed.

`xlim` GRAPH. Vector with the X axis limits.

`ylim` GRAPH. Vector with the Y axis limits.

`spacing` GRAPH. Space between the bars, both horizontal and vertical, which can be extended by changing the value "sp=0" to a larger number. See the function spacings for more options.

`keep_aspect_ratio`
GRAPH. Logical value, if TRUE the height and width of the graph are the same.

`xscale` GRAPH. The categories bar width.

`yspace` GRAPH. Vertical spacing between bars that can be expressed in different units as "lines", "cm", "mm", "inches", etc. See funtion unit for more details.

`main` GRAPH. Main title of the graph.

`sub` GRAPH. Subtitle of the graph.

`residuals_type` GRAPH. It specifies the type of residuals but, for the moment, this only allows the Pearson residuals.

`shade` GRAPH. Logical value that if TRUE are the results of the statistical Chi square of Pearson and, in addition, the categories that are significantly different are shaded.

`gp_axis` GRAPH. It defines the line type, thickness, color, etc., of the line of the bars of the categories.

`grid.edit` GRAPH. Logical value that if TRUE allows to highlight a category with a different color.

`text.grid` GRAPH. Text that allows to identify the variables and categories of variables to highlight, for example "rect:Parents=Yes,Sex=Male". If text.grid is FALSE, it is not used. See one of the examples for more details.

`gp.grid` GRAPH. Object of the class gpar which allows to define the transparency and the color of the category to highlight.

`labeling` GRAPH. It allows to set the labels for the categories on the left with "labeling_left", a framework with "labeling_cboxed", change from left to right and from top to bottom and put a framework with "labeling_lboxed", etc. See function labeling_border for more details.

`labeling_args` GRAPH. It allows to specify the distance at which the category labels, legends, etc., are placed. See one of the examples for more details.

`margins` GRAPH. Object of type unit with four components that delimit the margins from the top, right, down and left on the chart. See one of the examples for more details.

`legend_width` GRAPH. Margin of the legend to the right edge.

`gp_varnames` GRAPH. Size, format and font type letter of the legends of variables.

`gp_labels` GRAPH. Size, format and font type letter of each of the categories of variables.

`main_gp` GRAPH. Size, format, and font type letter of the main title.

`sub_gp` GRAPH. Size, format, and font type letter of the subtitle.

`legend` GRAPH. It allows to modify many aspects of the legend as the font size, font type, number of decimal places, change the text, etc.

`gp` GRAPH. It allows to change the scale of color in the legend and also has different palettes: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" and "shading_binary".

`file` TXT FILE. Output file name.

Details

VII. CONTRASTS OF HOMOGENEITY IN QUALITATIVE VARIABLES

In this function, the statistical treatments are performed for the contrasts of homogeneity that serve to work with qualitative variables, which are variables that are not expressed numerically and may be dichotomous (the variable can only take two values, such as women or man) or polytomous (the variable can have three or more values), being within these last two types: nominal and ordinal.

In a contrast of homogeneity, the aim is to check if two or more samples belong to the same population or that two samples belonging to the same population have not been altered. In other words, it would be interesting to know if the data from the measurements of a variable in several groups or times have the same distribution, or whether each measurement is different. Two types of contrasts depending on the nature of the variables can be distinguished, since it can work with independent measurements (for example, two selected samples in different areas) or paired (for example, the same variable can be measured twice).

VII.1. INDEPENDENT SAMPLES

VII.1.1. Polytomous variables

The tests are the Chi-square and the likelihood ratio (G-test).

VII.1.1.1. Chi-square

The null hypothesis of this test assumes that the samples belong to the same population and, therefore, the proportion of expected frequencies is the same in all samples. That can be used to calculate the expected frequency values for each cell, by multiplying the two marginal frequencies and dividing by the sample size, and are compared with the observed frequencies. In the event that the differences are small and, in consequence, the statistic of contrast lower than a critical value, the hypothesis that the samples are homogeneous (they have the same distribution) is accepted.

The formula for calculating the expected values is::

$$E_{ij} = \frac{n_{mi} n_{cj}}{n}$$

Once obtained the expected values, the statistic of contrast χ^2 is calculated with the following formula:

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

The critical value for (*h-1*)*(*k-1*) degrees of freedom and for the chosen level of significance (usually 0.05) is searched in Table χ^2 (Table 3 of Appendix I in Guisande et al., 2011). If χ^2 critical is greater than χ^2 calculated, the null hypothesis that are homogeneous is accepted, while if χ^2 critical is less than χ^2 calculated, (or p < significance level), the hypothesis is rejected. For the implementation of this test it is required that the expected frequencies do not take values lower than 5.

VII.1.1.2 Likelihood ratio (test G)

This contrast is very similar to test χ^2, since it also quantifies differences between observed and expected values. In addition, the degrees of freedom are calculated in the same way as in the test χ^2 of homogeneity of samples and the critical value of the test is obtained from the Table χ^2 (Table 3 of Appendix I in Guisande et al., 2011).

The contingency table and the calculation of the expected values are the same as those used in the χ^2 test of homogeneity.

The formula for the test statistic is:

$$G = 2\sum_{m=1}^{h}\sum_{c=1}^{k} O_{mc} ln \frac{O_{mc}}{E_{mc}}$$

Same as in the previous contrast, the critical value with the degrees of freedom and the level of certain significance is searched in Table χ^2 (Table 3 of Appendix I in Guisande et al., 2011). If χ^2 critical is greater than G, the null hypothesis that the samples are homogeneous is accepted.

The statistical G has a χ^2 distribution approximately. There are some corrections that improve the approximation, such as Williams (1976), which are applicable when the different categories have similar frequencies.

FUNCTIONS

The function assoc to make the graph (Meyer et al., 2006; 2013) was used. For more details on how to use this function, refer to the function help reference and/or Guisande & Vaamonde (2012).

EXAMPLE 1

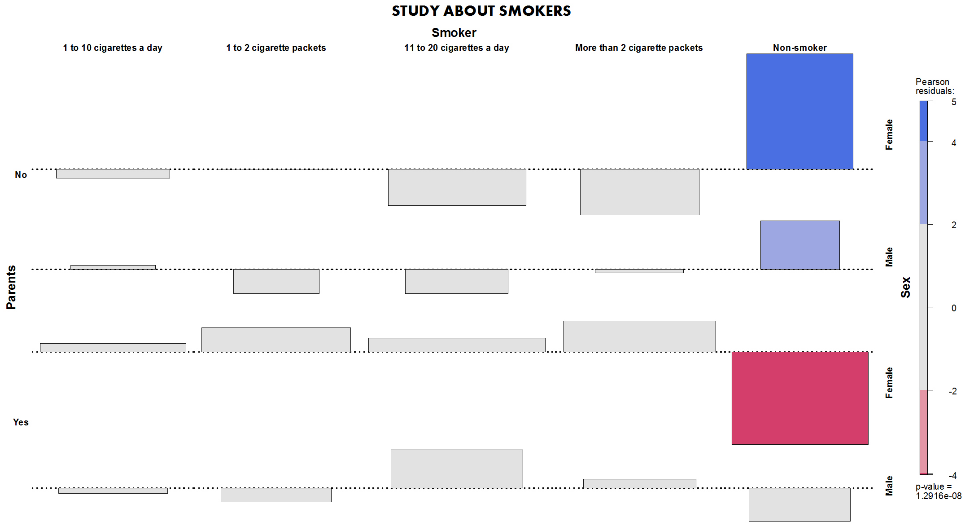
Data on the extent to which men and women smoke in different work centers. The categories used were: 1 (non-smoker), 2 (1 to 10 cigarettes a day), 3 (11 to 20 cigarettes a day), 4 (1 to 2 packs a day) and 5 (more than 2 packs a day).

There is also information about whether any of the parents of these workers are smokers and the categories are: workers in which either of their parents is smoker (category with value YES) and the other group are those in which none of their parents is smoker (category with value NO). It is a question of determining, in men and women, if there are differences between the group of people with any of the smoking parents and the group that their parents do not smoke, in the proportion of different types of degrees of smoker. For this, the variables to compare with the argument *variables = c("Parents", "Smoker")* are defined in the script and as it is intended that the analysis is done by separating sex, the argument is defined *group="Sex"*.

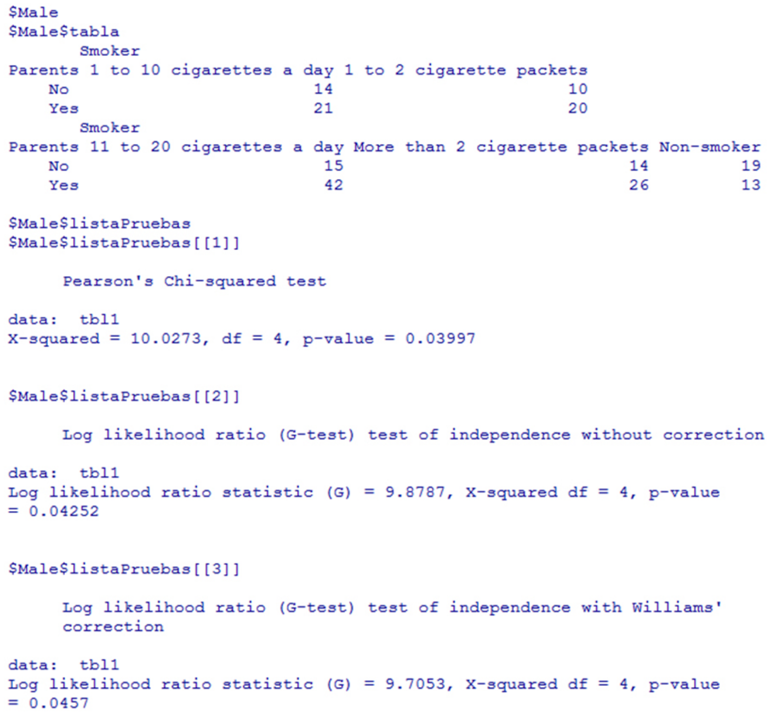
Figure VII1.1 shows the graphical representation of the contingency table. This chart also shows the results of the test χ^2 with a $p < 0.001$. Therefore, the null hypothesis that the samples are homogeneous is rejected, that is to say, there are significant differences between the group with one of their smoking parents and the group without smoking parents, in the percentage of the different types of smokers. The chart also gives a very important additional information, the categories where the differences are significant and which are contributing to that, as a whole, the test χ^2 is significant.

The gray color categories are not significantly different, while the categories with color are significantly different. In particular, it notes that under the category of non-smoking parents, for both men and women, there is a significantly higher proportion of people who do not smoke (the bars are above the dotted line). On the contrary, the number of people who do not smoke is significantly lower in the group of smoking parents (the bar is below the dotted line). Therefore, the fact that the parents do not smoke appears to foster their children not to smoke. However, once a person smokes, the degree to which smoke does not vary depending on whether the parents smoke or not smoke, as it could be observed that in all the categories of smokers, the color of the bars is gray.

Figure VII1.1 Graphical representation of the contingency table of variables sex, if the parents smoke or not and the level of smoking dependence.



```
$Female
$Female$tabla
       Smoker
Parents 1 to 10 cigarettes a day 1 to 2 cigarette packets
    No                        22                       25
    Yes                       42                       48
       Smoker
Parents 11 to 20 cigarettes a day More than 2 cigarette packets Non-smoker
    No                         26                            16         44
    Yes                        63                            52         11

$Female$listaPruebas
$Female$listaPruebas[[1]]

     Pearson's Chi-squared test

data:  tbl1
X-squared = 50.8757, df = 4, p-value = 2.37e-10


$Female$listaPruebas[[2]]

     Log likelihood ratio (G-test) test of independence without correction

data:  tbl1
Log likelihood ratio statistic (G) = 50.9257, X-squared df = 4, p-value
= 2.313e-10


$Female$listaPruebas[[3]]

     Log likelihood ratio (G-test) test of independence with Williams'
     correction

data:  tbl1
Log likelihood ratio statistic (G) = 50.4451, X-squared df = 4, p-value
= 2.915e-10
```



```
$Male
$Male$tabla
       Smoker
Parents 1 to 10 cigarettes a day 1 to 2 cigarette packets
    No                        14                       10
    Yes                       21                       20
       Smoker
Parents 11 to 20 cigarettes a day More than 2 cigarette packets Non-smoker
    No                         15                            14         19
    Yes                        42                            26         13

$Male$listaPruebas
$Male$listaPruebas[[1]]

     Pearson's Chi-squared test

data:  tbl1
X-squared = 10.0273, df = 4, p-value = 0.03997


$Male$listaPruebas[[2]]

     Log likelihood ratio (G-test) test of independence without correction

data:  tbl1
Log likelihood ratio statistic (G) = 9.8787, X-squared df = 4, p-value
= 0.04252


$Male$listaPruebas[[3]]

     Log likelihood ratio (G-test) test of independence with Williams'
     correction

data:  tbl1
Log likelihood ratio statistic (G) = 9.7053, X-squared df = 4, p-value
= 0.0457
```

In the script the argument was defined *group="Sex"* and, therefore, the test for men and women were carried out separately. In the case of men, both the test χ^2 (p = 0.039) as the test of the *G* with Williams' correction (p = 0.0457) showed significant differences. Therefore, there are significant differences between the group with one of their smoking parents and the group without parental smoking in the percentage of different types of smoking in men, which as shown in Figure VII1.1, were due to a more non-smoking men in the group of non-smoking parents.

In the case of women, the differences were even more significant, both in the test χ^2 ($p < 0.001$) as in the test of the *G* with the Williams' correction ($p < 0.001$). As well as men, the figure VII1.1 showed that these differences were due to a greater number of non-smoking women in the group of non-smoking parents.

EXAMPLE 2

In this second example the objective, using the same data for the previous example, is to compare whether the proportion of men and women sampled in the study was different in the group of smokers compared with non-smokers.

Figure VII1.2 Graphical representation of the contingency table of variables sex and if the parents smoke or not.

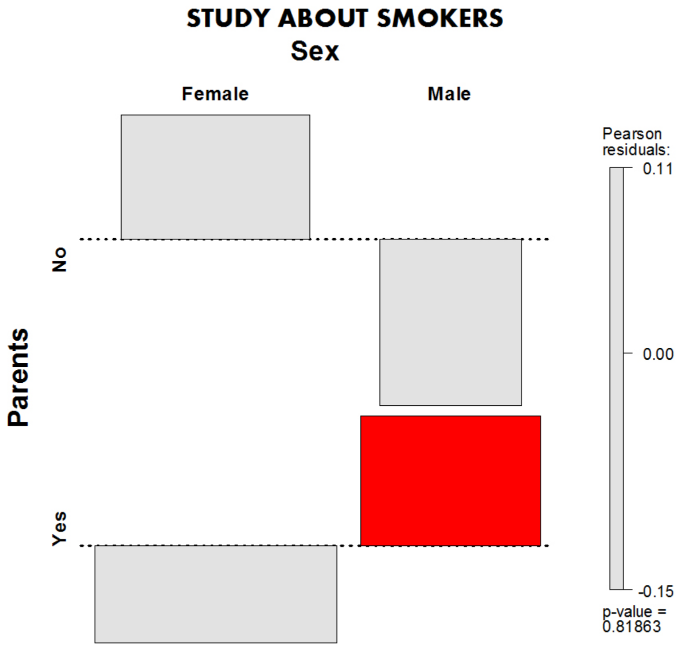


Figure VII1.2 shows no significant differences(χ^2, p = 0.818). Therefore, the proportion men and women was equal in the group of smoking and non-smoking parents. The example also shows how to highlight in a different color a category with the argument *text.grid="rect:Parents=Yes,Sex=Male"* (Figure VII1.2).

Value

A TXT file with the results of the contrast is obtained and the graph that shows the contingency table, if the option to display it was selected.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Meyer, D., Zeileis, A. & Hornik, K. (2006) The strucplot framework: Visualizing multiway contingency tables with vcd. *Journal of Statistical Software*, 17: 1-48.

Meyer, D., Zeileis, A. & Hornik, K. (2013) Visualizing Categorical Data. R package version 1.3-1. Available at: http://CRAN.R-project.org/package=vcd.

Williams, D.A. (1976) Improved llikelihood ratio test for complete contingency tables. *Biometrika*, 63: 33-37.

Examples

```
## Not run:
```

```
data(ZVII1)

#Example 1

#Differences in the degree of smoking depending on whether
#parents are smokers, and separating them by sex

VII1(data=ZVII1, variables=c("Parents","Smoker"), group="Sex",
test=c(1,2,3), labeling_args= list(offset_varnames = c(left =-0.5, top=-0.5),
offset_labels=c(top=-0.3), rot_labels=c(left=0)), main="STUDY ABOUT SMOKERS",
main_gp=gpar(fontsize=20, fontface=2, fontfamily="Aharoni"),
gp_varnames = gpar(fontsize=16,fontface=2),
gp_labels=gpar(fontsize = 12, fontface = 2))

#Example 2

#Shows how to highlight a category

VII1(data=ZVII1, variables=c("Parents","Sex"), test=c(1,2),
grid.edit=TRUE,
text.grid="rect:Parents=Yes,Sex=Male", gp.grid = gpar(fill = "red"),
margins=unit(c(4,2,2,4), "lines"), main="STUDY ABOUT SMOKERS", main_gp=gpar(fontsize=20,
fontface=2, fontfamily="Aharoni"), gp_varnames = gpar(fontsize=20, fontface=2),
gp_labels=gpar(fontsize = 14, fontface = 2))


## End(Not run)
```

VII2 *CONTRASTS OF HOMOGENEITY IN QUALITATIVE VARIABLES - INDEPENDENT SAMPLES AND DICHOTOMOUS VARIABLES*

Description

It determines if multiple samples, taken from the same population or of different populations, differ in a certain qualitative variable.

Usage

```
VII2(data, variables, group=NULL, test=c(1,4,5,6,7), graph=TRUE,
compress=TRUE, xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

`data` Data file.

`variables` Variable or variables to which the contrast of homogeneity is to be performed.

`group` Variables together with the data are grouped for calculations. In case of selecting NULL, no grouping would be done, and this would be calculated considering all data of the selected variables.

`test` TYPE OF CONTRAST OF HOMOGENEITY.

1. Chi-square test.
2. G-test without the Williams' correction.
3. G-test with the Williams' correction.
4. Chi-square test with Yates correction (Only for 2x2 tables).
5. Fisher's exact test (Only for 2x2 tables), bilateral.
6. Fisher's exact test (Only for 2x2 tables), unilateral: less than.
7. Fisher's exact test (Only for 2x2 tables), unilateral: greater than.
8. Friedman ANOVA test.
9. McNemar's test without the continuity correction.
10. McNemar's test with the continuity correction (Only for 2x2 tables).
11. Cochran test.

`graph` GRAPH. If TRUE, the graph of the contingency table is shown. This chart does not allow accents and the categories are sorted by alphabetical order. The example of the function VIII1 shows how to sort the categories.

`compress` GRAPH. Logical value, if FALSE the space between rows and columns is chosen so that the total of heights and widths of the rows and columns are equal. If TRUE, the space between rows and columns is fixed and, therefore, the graph is more compressed.

`xlim` GRAPH. Vector with the X axis limits.

`ylim` GRAPH. Vector with the Y axis limits.

`spacing` GRAPH. Space between the bars, both horizontal and vertical, which can be extended by changing the value "sp=0" to a larger number. See the function spacings for more options.

`keep_aspect_ratio`
GRAPH. Logical value, if TRUE the height and width of the graph are the same.

`xscale` GRAPH. The categories bar width.

`yspace` GRAPH. Vertical spacing between bars that can be expressed in different units as "lines", "cm", "mm", "inches", etc. See funtion unit for more details.

`main` GRAPH. Main title of the graph.

`sub` GRAPH. Subtitle of the graph.

`residuals_type` GRAPH. It specifies the type of residuals but, for the moment, this only allows the Pearson residuals.

`shade` GRAPH. Logical value that if TRUE are the results of the statistical Chi square of Pearson and, in addition, the categories that are significantly different are shaded.

`gp_axis` GRAPH. It defines the line type, thickness, color, etc., of the line of the bars of the categories.

`grid.edit` GRAPH. Logical value that if TRUE allows to highlight a category with a different color.

`text.grid` GRAPH. Text that allows to identify the variables and categories of variables to highlight, for example "rect:Parents=Yes,Sex=Male". If text.grid is FALSE, it is not used. See one of the examples for more details.

`gp.grid` GRAPH. Object of the class gpar which allows to define the transparency and the color of the category to highlight.

`labeling` GRAPH. It allows to set the labels for the categories on the left with "labeling_left", a framework with "labeling_cboxed", change from left to right and from top to bottom and put a framework with "labeling_lboxed", etc. See function labeling_border for more details.

`labeling_args` GRAPH. It allows to specify the distance at which the category labels, legends, etc., are placed. See one of the examples for more details.

`margins` GRAPH. Object of type unit with four components that delimit the margins from the top, right, down and left on the chart. See one of the examples for more details.

`legend_width` GRAPH. Margin of the legend to the right edge.

`gp_varnames` GRAPH. Size, format and font type letter of the legends of variables.

`gp_labels` GRAPH. Size, format and font type letter of each of the categories of variables.

`main_gp` GRAPH. Size, format, and font type letter of the main title.

`sub_gp` GRAPH. Size, format, and font type letter of the subtitle.

`legend` GRAPH. It allows to modify many aspects of the legend as the font size, font type, number of decimal places, change the text, etc.

`gp` GRAPH. It allows to change the scale of color in the legend and also has different palettes: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" and "shading_binary".

`file` TXT FILE. Output file name.

Details

VII. CONTRASTS OF HOMOGENEITY IN QUALITATIVE VARIABLES

VII.1. INDEPENDENT SAMPLES

VII.1.2. Dichotomous variables

When both variables are dichotomous, i.e. that have only two categories such as, for example, presence/absence or woman/man, then a contingency table of 2x2 is obtained and a correction of the test χ^2 is used, the Yates' correction (1934) or the Fisher's exact test (1922).

VII.1.2.1. The Yates' correction

The Yates' correction is applied when a contingency table is 2x2, i.e., two rows and two columns, and the sample has a small number of cases since it is not possible to group categories.

In this situation, the value of χ^2 is overestimated and, to correct it, the so-called Yates correction that modifies the calculation of the statistical is applied as follows:

$$\chi^2 = \sum_{m=1}^{j} \sum_{c=1}^{i} \frac{|O_{mc} - E_{mcj}| - \frac{1}{2})^2}{E_{mc}}$$

In this case the critical value of contrast would be the χ^2 for a degree of freedom. If the value obtained is less than the tabulated value of contrast, the null hypothesis that the samples are homogeneous is accepted.

However, it is worth mentioning that the use of the Yates' correction does not exempt from certain requirements about the sample size necessary for the use of statistical χ^2. As a general rule, it will be required that 80% of the cells in a contingency table have expected values greater than 5. Thus, in a 2x2 table all of the cells will have to verify this condition, although in practice is allowed one of them to display expected frequencies slightly below this value.

VII.1.2.2. Fisher's test

Fisher's exact test is used with contingency tables 2x2 and is the most advisable when there is one or more cells with expected frequencies lower than 5.

The test is based on the determination of the probability of obtaining the observed frequencies, from a hypergeometric distribution, when the total of rows and columns are constant, taking as null hypothesis that the variables are independent.

FUNCTIONS

The function assoc to make the graph (Meyer et al., 2006; 2013) was used. For more details on how to use this function, refer to the function help reference and/or Guisande & Vaamonde (2012).

EXAMPLE

Data for the presence of the virus of herpes labialis in men and women. The aim is to determine if the number of carriers of the virus is different between men and women. This specifies the variables to compare in the argument *variables = c("Sex", "Presence").*

Figure VII2.1 Graphical representation of the contingency table of the variables sex and presence of the virus of herpes labialis.

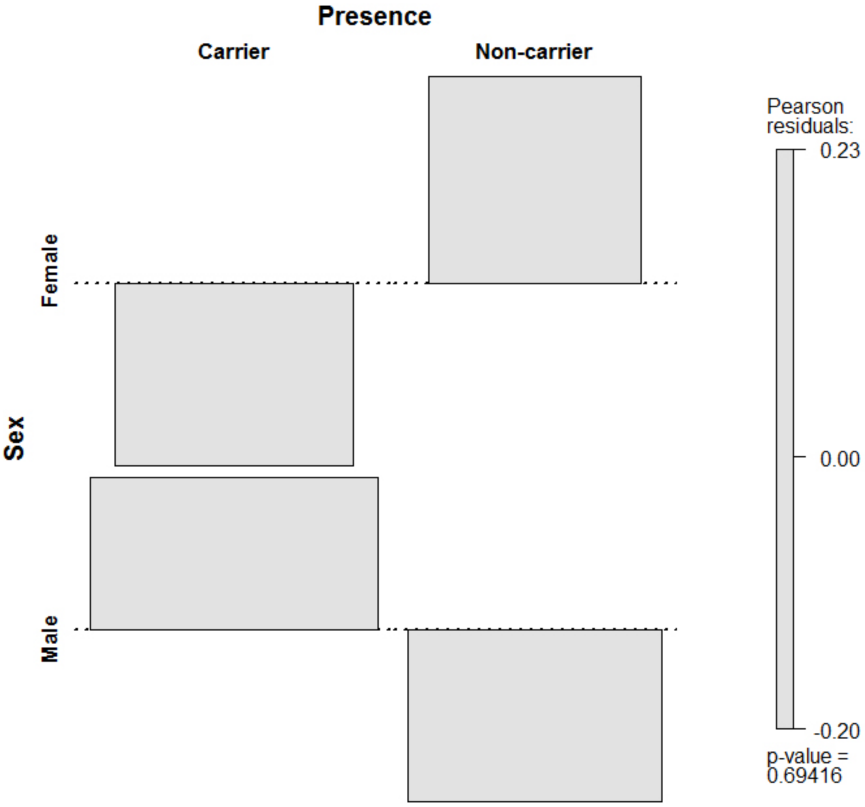


Figure VII2.1 shows that there are no significant differences with p = 0.694. Therefore, the presence of the virus of herpes labialis is equal in men and women.

The results of all tests are shown below:

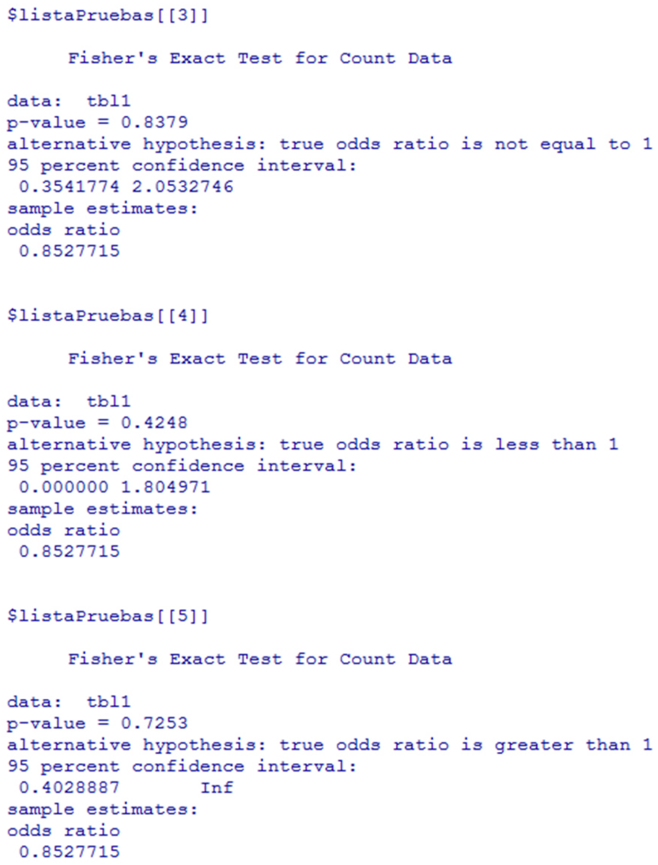
Chi-Square

The value χ^2 is 0.15 with p = 0.694. It is therefore concluded that there was no relationship between gender and the presence of the virus of herpes labialis.

Fisher's exact test.

It returns the calculated statistical both with a unilateral alternative hypothesis (that tells whether the number of females who are carriers of the virus is greater than the number of men carriers) as a two-sided alternative hypothesis (which simply reports whether the virus is different).

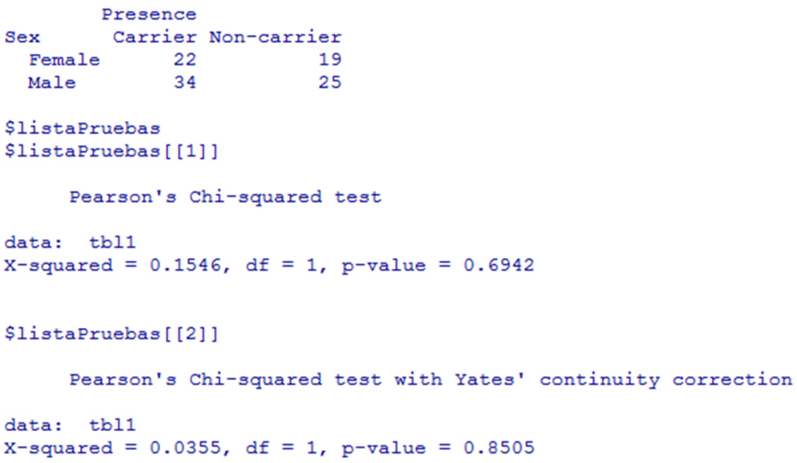
The odds would be unilateral p = 0.425 and p = 0.838 for bilateral. In both cases, the assumption of homogeneity of variables is accepted, although with these data (it is not an a priori assumption that there will be more women with virus than men), the bilateral probability should be applied.



```
$listaPruebas[[3]]

     Fisher's Exact Test for Count Data

data:  tbl1
p-value = 0.8379
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3541774 2.0532746
sample estimates:
odds ratio
 0.8527715


$listaPruebas[[4]]

     Fisher's Exact Test for Count Data

data:  tbl1
p-value = 0.4248
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 1.804971
sample estimates:
odds ratio
 0.8527715


$listaPruebas[[5]]

     Fisher's Exact Test for Count Data

data:  tbl1
p-value = 0.7253
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.4028887       Inf
sample estimates:
odds ratio
 0.8527715
```

Yates' correction.

In this case, and given that the frequencies of each category are greater than 5, this correction would not be necessary. Assuming that it would need, a contrast value of 0.035 and p = 0.851, so the assumption that the presence of the virus is the same in both sexes is accepted.



```
        Presence
Sex      Carrier Non-carrier
  Female      22          19
  Male        34          25

$listaPruebas
$listaPruebas[[1]]

     Pearson's Chi-squared test

data:  tbl1
X-squared = 0.1546, df = 1, p-value = 0.6942


$listaPruebas[[2]]

     Pearson's Chi-squared test with Yates' continuity correction

data:  tbl1
X-squared = 0.0355, df = 1, p-value = 0.8505
```

Value

A TXT file is obtained with the results of the contrasts and the graph that shows the contingency table, if the option of displaying it is selected.

References

Fisher, L.D. & van Belle, G. (1993) *Biostatistics. A methodology for the health sciences*. John Wiley y Sons, New York.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Meyer, D., Zeileis, A. & Hornik, K. (2006) The strucplot framework: Visualizing multiway contingency tables with vcd. *Journal of Statistical Software*, 17: 1-48.

Meyer, D., Zeileis, A. & Hornik, K. (2013) Visualizing Categorical Data. R package version 1.3-1. Available at: http://CRAN.R-project.org/package=vcd.

Yates, F. (1934) Contingency table involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1: 217-235.

Examples

```
## Not run:

data(ZVII2)

#Presence of the herpes virus

VII2(data = ZVII2, variables = c("Sex","Presence"), gp_varnames = gpar(fontsize=15,
fontface=2), gp_labels=gpar(fontsize = 13, fontface = 2))


## End(Not run)
```

VII3 *CONTRASTS OF HOMOGENEITY IN QUALITATIVE VARIABLES-RELATED SAMPLES AND POLYTOMOUS VARIABLES*

Description

It determines if multiple samples, taken from the same population or of different populations, differ in a certain qualitative variable.

Usage

```
VII3(data, variables, group=NULL, test=c(8), graph=TRUE,
compress=TRUE,
xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

`data`	Data file.
`variables`	Variable or variables to which the contrast of homogeneity is to be performed.
`group`	Variables together with the data are grouped for calculations. In case of selecting NULL, no grouping would be done, and this would be calculated considering all data of the selected variables.
`test`	TYPE OF CONTRAST OF HOMOGENEITY. 1. Chi-square test. 2. G-test without the Williams' correction. 3. G-test with the Williams' correction. 4. Chi-square test with Yates correction (Only for 2x2 tables). 5. Fisher's exact test (Only for 2x2 tables), bilateral. 6. Fisher's exact test (Only for 2x2 tables), unilateral: less than. 7. Fisher's exact test (Only for 2x2 tables), unilateral: greater than. 8. Friedman ANOVA test. 9. McNemar's test without the continuity correction. 10. McNemar's test with the continuity correction (Only for 2x2 tables). 11. Cochran test.

`graph` GRAPH. If TRUE, the graph of the contingency table is shown. This chart does not allow accents and the categories are sorted by alphabetical order. The example of the function VIII1 shows how to sort the categories.

`compress` GRAPH. Logical value, if FALSE the space between rows and columns is chosen so that the total of heights and widths of the rows and columns are equal. If TRUE, the space between rows and columns is fixed and, therefore, the graph is more compressed.

`xlim` GRAPH. Vector with the X axis limits.

`ylim` GRAPH. Vector with the Y axis limits.

`spacing` GRAPH. Space between the bars, both horizontal and vertical, which can be extended by changing the value "sp=0" to a larger number. See the function spacings for more options.

`keep_aspect_ratio`
GRAPH. Logical value, if TRUE the height and width of the graph are the same.

`xscale` GRAPH. The categories bar width.

`yspace` GRAPH. Vertical spacing between bars that can be expressed in different units as "lines", "cm", "mm", "inches", etc. See funtion unit for more details.

`main` GRAPH. Main title of the graph.

`sub` GRAPH. Subtitle of the graph.

`residuals_type` GRAPH. It specifies the type of residuals but, for the moment, this only allows the Pearson residuals.

`shade` GRAPH. Logical value that if TRUE are the results of the statistical Chi square of Pearson and, in addition, the categories that are significantly different are shaded.

`gp_axis` GRAPH. It defines the line type, thickness, color, etc., of the line of the bars of the categories.

`grid.edit` GRAPH. Logical value that if TRUE allows to highlight a category with a different color.

`text.grid` GRAPH. Text that allows to identify the variables and categories of variables to highlight, for example "rect:Parents=Yes,Sex=Male". If text.grid is FALSE, it is not used. See one of the examples for more details.

`gp.grid` GRAPH. Object of the class gpar which allows to define the transparency and the color of the category to highlight.

`labeling` GRAPH. It allows to set the labels for the categories on the left with "labeling_left", a framework with "labeling_cboxed", change from left to right and from top to bottom and put a framework with "labeling_lboxed", etc. See function labeling_border for more details.

`labeling_args` GRAPH. It allows to specify the distance at which the category labels, legends, etc., are placed. See one of the examples for more details.

`margins` GRAPH. Object of type unit with four components that delimit the margins from the top, right, down and left on the chart. See one of the examples for more details.

`legend_width` GRAPH. Margin of the legend to the right edge.

`gp_varnames` GRAPH. Size, format and font type letter of the legends of variables.

`gp_labels` GRAPH. Size, format and font type letter of each of the categories of variables.

`main_gp` GRAPH. Size, format, and font type letter of the main title.

`sub_gp` GRAPH. Size, format, and font type letter of the subtitle.

`legend` GRAPH. It allows to modify many aspects of the legend as the font size, font type, number of decimal places, change the text, etc.

`gp` GRAPH. It allows to change the scale of color in the legend and also has different palettes: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" and "shading_binary".

`file` TXT FILE. Output file name.

Details

VII. CONTRASTS OF HOMOGENEITY IN QUALITATIVE VARIABLES

VII.2. RELATED SAMPLES

It may be the case that the variables to be compared are related, i.e., they are paired data. For example, when a treatment is given to a group of people and it is studied whether this drug causes some type of allergy, allergy is quantified by means of any qualitative variable as presence or absence of spots on the skin. In this case, the variables to compare are not independent, since they are the same individuals that are compared over time.

As in the case of the independent samples, there is evidence for polytomous and dichotomous variables.

VII.2.1. Polytomous variables

VII.2.1.1. The Friedman ANOVA

The test developed by Friedman (1937), is the similar test to analysis of variance (ANOVA) for repeated samples, despite the fact that it is a χ^2 contrast, it is called ANOVA.

The steps in this statistical are described by Siegel & Castellan (1988). They are as follows:

1. Assign ranges within each element in the sample.

2. Add the ranges of each element in the sample.

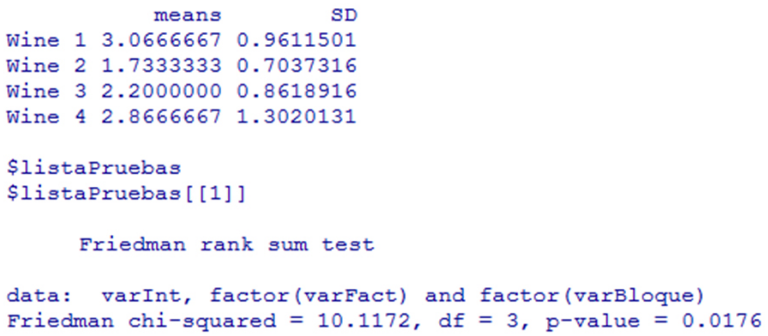
3. Calculate the statistic χ^2 to contrast and compare it with the tabulated value χ^2 critical with a - 1 degrees of freedom, being a the number of variables.

EXAMPLE

Data on how 15 enologists evaluated 4 different types of wines. Category 1 is Very bad, 2 Poor, 3 Good and 4 Very good. The objective is to determine if there are differences between the quality of the wines. As the evaluation, of each of the enologists of different wines, is compared, the data are not independent, they are related. In addition each of the variables has more than 2 values, so they are polytomous.

In the case of the Friedman ANOVA, the following order is necessary: the variable that has the data from the response, the factor, and, finally, the variable that has the information about the individuals to which each response belongs, in this case would be *variables=c("Quality","Wine","Oenologist")*.

Since the probability is less than 0.05, the null hypothesis of homogeneity is rejected and therefore, there are significant differences between the quality of wines according to the enologists (ANOVA χ^2 = 10.1, df = 3, p = 0.018).



```
             means        SD
Wine 1 3.0666667 0.9611501
Wine 2 1.7333333 0.7037316
Wine 3 2.2000000 0.8618916
Wine 4 2.8666667 1.3020131

$listaPruebas
$listaPruebas[[1]]

      Friedman rank sum test

data:  varInt, factor(varFact) and factor(varBloque)
Friedman chi-squared = 10.1172, df = 3, p-value = 0.0176
```

Value

A TXT file is obtained with the results of the contrasts and the graph that shows the contingency table, if the option of displaying it is selected.

References

Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32: 675-701.

Siegel, S. & Castellan, N.J. Jr. (1988) *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Examples

```
## Not run:

data(ZVII3)

#Wine-tasting

VII3(data=ZVII3, variables=c("Quality","Wine","Oenologist"),
graph=FALSE)


## End(Not run)
```

VII4 *CONTRASTS OF HOMOGENEITY IN QUALITATIVE VARIABLES-RELATED SAMPLES AND DICHOTOMOUS VARIABLES WITHOUT REPETITION*

Description

It determines if multiple samples, taken from the same population or from different populations, differ in a certain qualitative variable.

Usage

```
VII4(data, variables, group=NULL, test=c(9,10), graph=TRUE, compress=TRUE,
xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

`data`	Data file.
`variables`	Variable or variables to which the contrast of homogeneity is to be performed.
`group`	Variables together with the data are grouped for calculations. In case of selecting NULL, no grouping would be done, and this would be calculated considering all data of the selected variables.
`test`	TYPE OF CONTRAST OF HOMOGENEITY. 1. Chi-square test. 2. G-test without the Williams' correction. 3. G-test with the Williams' correction. 4. Chi-square test with Yates correction (Only for 2x2 tables). 5. Fisher's exact test (Only for 2x2 tables), bilateral. 6. Fisher's exact test (Only for 2x2 tables), unilateral: less than. 7. Fisher's exact test (Only for 2x2 tables), unilateral: greater than. 8. Friedman ANOVA test. 9. McNemar's test without the continuity correction. 10. McNemar's test with the continuity correction (Only for 2x2 tables). 11. Cochran test.
`graph`	GRAPH. If TRUE, the graph of the contingency table is shown. This chart does not allow accents and the categories are sorted by alphabetical order. The example of the function VIII1 shows how to sort the categories.
`compress`	GRAPH. Logical value, if FALSE the space between rows and columns is chosen so that the total of heights and widths of the rows and columns are equal. If TRUE, the space between rows and columns is fixed and, therefore, the graph is more compressed.
`xlim`	GRAPH. Vector with the X axis limits.
`ylim`	GRAPH. Vector with the Y axis limits.

`spacing` GRAPH. Space between the bars, both horizontal and vertical, which can be extended by changing the value "sp=0" to a larger number. See the function spacings for more options.

`keep_aspect_ratio`
GRAPH. Logical value, if TRUE the height and width of the graph are the same.

`xscale` GRAPH. The categories bar width.

`yspace` GRAPH. Vertical spacing between bars that can be expressed in different units as "lines", "cm", "mm", "inches", etc. See funtion unit for more details.

`main` GRAPH. Main title of the graph.

`sub` GRAPH. Subtitle of the graph.

`residuals_type` GRAPH. It specifies the type of residuals but, for the moment, this only allows the Pearson residuals.

`shade` GRAPH. Logical value that if TRUE are the results of the statistical Chi square of Pearson and, in addition, the categories that are significantly different are shaded.

`gp_axis` GRAPH. It defines the line type, thickness, color, etc., of the line of the bars of the categories.

`grid.edit` GRAPH. Logical value that if TRUE allows to highlight a category with a different color.

`text.grid` GRAPH. Text that allows to identify the variables and categories of variables to highlight, for example "rect:Parents=Yes,Sex=Male". If text.grid is FALSE, it is not used. See one of the examples for more details.

`gp.grid` GRAPH. Object of the class gpar which allows to define the transparency and the color of the category to highlight.

`labeling` GRAPH. It allows to set the labels for the categories on the left with "labeling_left", a framework with "labeling_cboxed", change from left to right and from top to bottom and put a framework with "labeling_lboxed", etc. See function labeling_border for more details.

`labeling_args` GRAPH. It allows to specify the distance at which the category labels, legends, etc., are placed. See one of the examples for more details.

`margins` GRAPH. Object of type unit with four components that delimit the margins from the top, right, down and left on the chart. See one of the examples for more details.

`legend_width` GRAPH. Margin of the legend to the right edge.

`gp_varnames` GRAPH. Size, format and font type letter of the legends of variables.

`gp_labels` GRAPH. Size, format and font type letter of each of the categories of variables.

`main_gp` GRAPH. Size, format, and font type letter of the main title.

`sub_gp` GRAPH. Size, format, and font type letter of the subtitle.

`legend` GRAPH. It allows to modify many aspects of the legend as the font size, font type, number of decimal places, change the text, etc.

`gp` GRAPH. It allows to change the scale of color in the legend and also has different palettes: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" and "shading_binary".

`file` TXT FILE. Output file name.

Details

VII. CONTRASTS OF HOMOGENEITY IN QUALITATIVE VARIABLES

VII.2. RELATED SAMPLES

VII.2.2. Dichotomous variable without repetition

VII.2.2.1. McNemar's test

It is used with samples that measured the same dichotomous variable twice, i.e., they are paired data (McNemar, 1947). It is widely used in medicine to see if a treatment has effect, being able to represent data in a 2x2 contingency table.

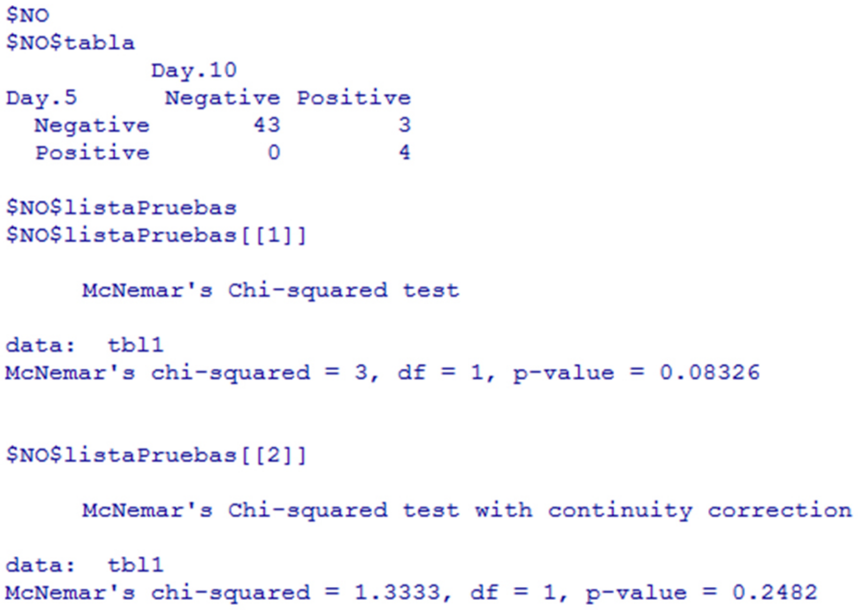
It performs the following contrast, which assumes as null hypothesis that the expected frequencies b and c are the same:

$$\chi^2 = \frac{(b-c)^2}{(b+c)}$$

The critical value of contrast corresponds to the distribution χ^2 to 1 degree of freedom.

EXAMPLE

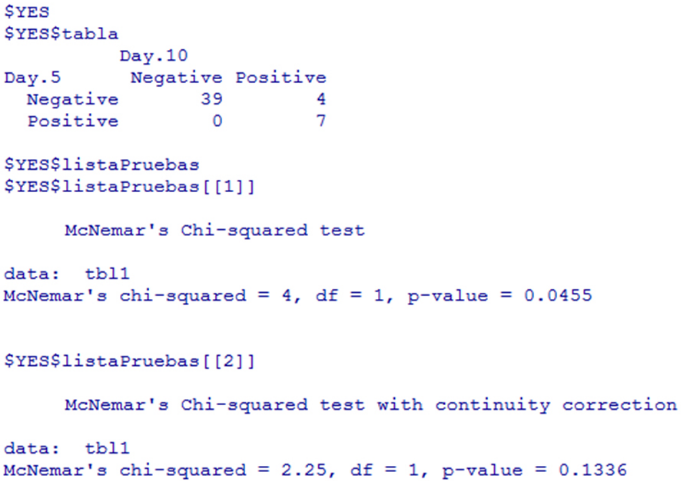
The objective of this study is to determine if an antibiotic causes an allergic response in the skin. For this reason the antibiotic was administered to 50 people and 50 others were given a placebo. The results compare positive response (allergy) or negative (no allergy) in 100 people between days 5 and 10 after treatment. What is tested is whether there are differences in each of the groups, that is, if they are homogeneous in terms of the allergic response, in the two periods considered, 5 and 10 days. Assuming that 5 days is before and 10 days is after and allergy necessarily appear in this interval, the allergic effect of the antibiotic can be tested if homogeneity was found in the placebo group and non-homogeneity in the group that was provided antibiotic, although in the latter case should also be checked that the allergy is higher after than before and not the other way around (in both cases, the contrast may be significant).



```
$NO
$NO$tabla
          Day.10
Day.5       Negative Positive
  Negative        43        3
  Positive         0        4

$NO$listaPruebas
$NO$listaPruebas[[1]]

     McNemar's Chi-squared test

data:  tbl1
McNemar's chi-squared = 3, df = 1, p-value = 0.08326


$NO$listaPruebas[[2]]

     McNemar's Chi-squared test with continuity correction

data:  tbl1
McNemar's chi-squared = 1.3333, df = 1, p-value = 0.2482
```

In the case of people who were not provided with the antibiotic, such as p = 0.248 in the test with continuity correction (which is the test that must be used when there are frequencies lower than 5), it is accepted the null hypothesis that there are no significant differences in the allergic response to the antibiotic between day 5 and day 10.

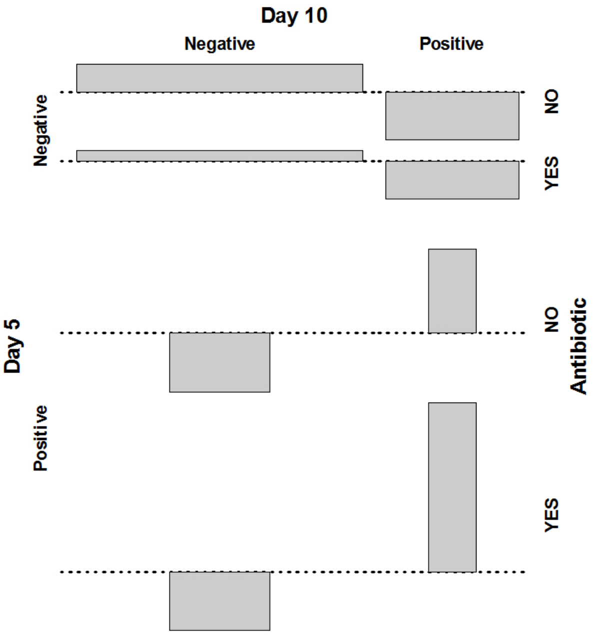
In the case of people who were provided with the antibiotic such as p = 0.134, to be greater than 0.05, it is also accepted the null hypothesis that there are no significant differences in the allergic response to the antibiotic between day 5 and day 10.



```
$YES
$YES$tabla
          Day.10
Day.5      Negative Positive
  Negative       39        4
  Positive        0        7

$YES$listaPruebas
$YES$listaPruebas[[1]]

     McNemar's Chi-squared test

data:  tbl1
McNemar's chi-squared = 4, df = 1, p-value = 0.0455


$YES$listaPruebas[[2]]

     McNemar's Chi-squared test with continuity correction

data:  tbl1
McNemar's chi-squared = 2.25, df = 1, p-value = 0.1336
```

Figure VII4.1 shows the results. Not the scale is shown with the significance because the test χ^2 is not used.

Figure VII4.1 Contingency table of the variables antibiotic and the days 5 and 10.



Value

A TXT file is obtained with the results of the contrasts and the graph that shows the contingency table, if the option of displaying it is selected.

References

McNemar Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12: 153-157.

Examples

```
## Not run:

data(ZVII4)

VII4(data=ZVII4, variables=c("Day.5", "Day.10"), group="Antibiotic",
shade=FALSE, labeling_args =list(set_varnames= c(Day.5="Day 5", Day.10="Day 10")),
gp_varnames = gpar(fontsize=15,fontface=2), gp_labels=gpar(fontsize=13, fontface=2))


## End(Not run)
```

VII5 *CONTRASTS OF HOMOGENEITY IN QUALITATIVE VARIABLES-RELATED SAMPLES AND DICHOTOMOUS VARIABLES WITH REPETITION*

Description

It determines if multiple samples, taken from the same population or from different populations, differ in a certain qualitative variable.

Usage

```
VII5(data, variables, group=NULL, test=c(11),graph=TRUE, compress=TRUE,
xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

`data`	Data file.
`variables`	Variable or variables to which the contrast of homogeneity is to be performed.
`group`	Variables together with the data are grouped for calculations. In case of selecting NULL, no grouping would be done, and this would be calculated considering all data of the selected variables.
`test`	TYPE OF CONTRAST OF HOMOGENEITY. 1. Chi-square test. 2. G-test without the Williams' correction. 3. G-test with the Williams' correction. 4. Chi-square test with Yates correction (Only for 2x2 tables). 5. Fisher's exact test (Only for 2x2 tables), bilateral. 6. Fisher's exact test (Only for 2x2 tables), unilateral: less than. 7. Fisher's exact test (Only for 2x2 tables), unilateral: greater than. 8. Friedman ANOVA test. 9. McNemar's test without the continuity correction. 10. McNemar's test with the continuity correction (Only for 2x2 tables). 11. Cochran test.

`graph` GRAPH. If TRUE, the graph of the contingency table is shown. This chart does not allow accents and the categories are sorted by alphabetical order. The example of the function VIII1 shows how to sort the categories.

`compress` GRAPH. Logical value, if FALSE the space between rows and columns is chosen so that the total of heights and widths of the rows and columns are equal. If TRUE, the space between rows and columns is fixed and, therefore, the graph is more compressed.

`xlim` GRAPH. Vector with the X axis limits.

`ylim` GRAPH. Vector with the Y axis limits.

`spacing` GRAPH. Space between the bars, both horizontal and vertical, which can be extended by changing the value "sp=0" to a larger number. See the function spacings for more options.

`keep_aspect_ratio`

GRAPH. Logical value, if TRUE the height and width of the graph are the same.

`xscale` GRAPH. The categories bar width.

`yspace` GRAPH. Vertical spacing between bars that can be expressed in different units as "lines", "cm", "mm", "inches", etc. See funtion unit for more details.

`main` GRAPH. Main title of the graph.

`sub` GRAPH. Subtitle of the graph.

`residuals_type` GRAPH. It specifies the type of residuals but, for the moment, this only allows the Pearson residuals.

`shade` GRAPH. Logical value that if TRUE are the results of the statistical Chi square of Pearson and, in addition, the categories that are significantly different are shaded.

`gp_axis` GRAPH. It defines the line type, thickness, color, etc., of the line of the bars of the categories.

`grid.edit` GRAPH. Logical value that if TRUE allows to highlight a category with a different color.

`text.grid` GRAPH. Text that allows to identify the variables and categories of variables to highlight, for example "rect:Parents=Yes,Sex=Male". If text.grid is FALSE, it is not used. See one of the examples for more details.

`gp.grid` GRAPH. Object of the class gpar which allows to define the transparency and the color of the category to highlight.

`labeling` GRAPH. It allows to set the labels for the categories on the left with "labeling_left", a framework with "labeling_cboxed", change from left to right and from top to bottom and put a framework with "labeling_lboxed", etc. See function labeling_border for more details.

`labeling_args` GRAPH. It allows to specify the distance at which the category labels, legends, etc., are placed. See one of the examples for more details.

`margins` GRAPH. Object of type unit with four components that delimit the margins from the top, right, down and left on the chart. See one of the examples for more details.

`legend_width` GRAPH. Margin of the legend to the right edge.

gp_varnames	GRAPH. Size, format and font type letter of the legends of variables.
gp_labels	GRAPH. Size, format and font type letter of each of the categories of variables.
main_gp	GRAPH. Size, format, and font type letter of the main title.
sub_gp	GRAPH. Size, format, and font type letter of the subtitle.
legend	GRAPH. It allows to modify many aspects of the legend as the font size, font type, number of decimal places, change the text, etc.
gp	GRAPH. It allows to change the scale of color in the legend and also has different palettes: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" and "shading_binary".
file	TXT FILE. Output file name.

Details

VII. CONTRASTS OF HOMOGENEITY IN QUALITATIVE VARIABLES

VII.2. RELATED SAMPLES

VII.2.3. Dichotomous variable with repetition

VII.2.3.1. Cochran's Q test

It is the equivalent of the McNemar test when measurements are performed more than twice. For the calculation of the statistical of contrast, the data is presented in a contingency table. The values of the contingency table must be transformed to 0 and 1. For example, if the variable was presence/absence of a character, it could indicate presence with 1 and absence with 0. The statistical is the following:

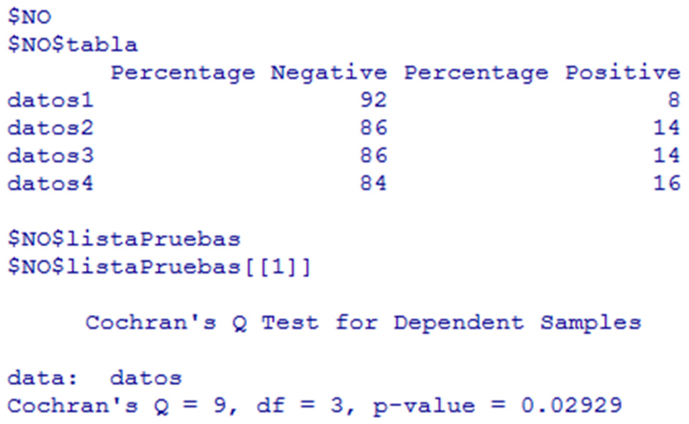
$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} c_j^2 - \left(\sum_{j=1}^{k} c_j\right)^2\right]}{k\sum_{j=1}^{k} c_j - \sum_{i=1}^{N} n_i^2}$$

Being *k* the number of dichotomous variables, and *N* the number of cases, c_j the number of successes for each variable, and n_i the number of successes for each case.

EXAMPLE

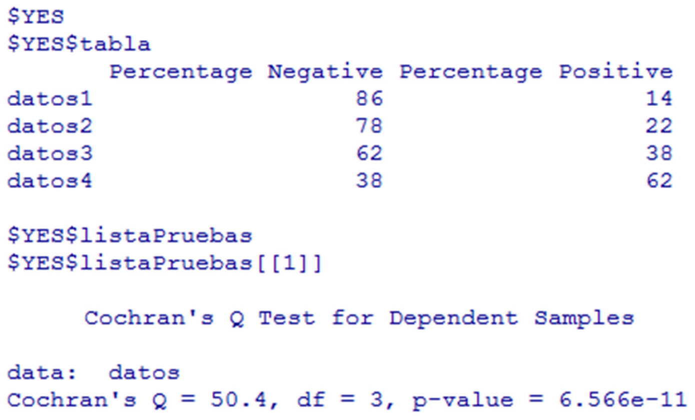
The objective of this study is to determine if an antibiotic causes an allergic response in the skin. For this reason the antibiotic was administered to 50 people and 50 others were given a placebo. The results compare positive response (allergy) or negative (no allergy) in 100 people in several days. What is tested is whether there are differences in each group, i.e., whether they are homogeneous in the allergic response, during the considered period. The allergic effect of the antibiotic can be tested if homogeneity was found in the placebo group and non-homogeneity in the group that was provided antibiotic, although in the latter case should also be checked that the allergy is higher after than before and not the other way around (in both cases, the contrast may be significant).

In the case of people who were not given the treatment, the probability is less than 0.05 so the null hypothesis that the samples are equal is rejected, that is to say, there are differences between the various days in allergy symptoms detected in people who are subjected to the placebo (Test *Q* = 9 of Cochran, df = 3, p = 0.03).



```
$NO
$NO$tabla
        Percentage Negative Percentage Positive
datos1                   92                   8
datos2                   86                  14
datos3                   86                  14
datos4                   84                  16

$NO$listaPruebas
$NO$listaPruebas[[1]]

      Cochran's Q Test for Dependent Samples

data:  datos
Cochran's Q = 9, df = 3, p-value = 0.02929
```

In the case of people who were given antibiotic, it also rejects the null hypothesis that the samples are equal, i.e., there are differences between different days in allergy symptoms detected (Test*Q*Cochran = 50.4, df = 3, p < 0.001).



```
$YES
$YES$tabla
        Percentage Negative Percentage Positive
datos1                   86                  14
datos2                   78                  22
datos3                   62                  38
datos4                   38                  62

$YES$listaPruebas
$YES$listaPruebas[[1]]

      Cochran's Q Test for Dependent Samples

data:  datos
Cochran's Q = 50.4, df = 3, p-value = 6.566e-11
```

It is important to highlight the fact that differences were observed over time in both groups (with and without antibiotic) means not definitively conclude that the antibiotic is not responsible for the allergy. This test only enables us to prove that there are differences in the allergic response over time. That is to say, it could be the case that in the people who were given placebo, the allergic response decreases with the time and, on the contrary, in those who received antibiotic increases. Therefore, it would be necessary to perform a higher contrast of independence as explained in the function VIII1, in order to be able to rule out the relation between the antibiotic and the allergy.

Value

A TXT file is obtained with the results of the contrasts and the graph that shows the contingency table, if the option of displaying it is selected.

Examples

```
## Not run:
```

```
data(ZVII5)

#Skin allergy to an antibiotic in several days

VII5(data=ZVII5, variables=c("Day.5","Day.10","Day.15","Day.20"), group="Antibiotic",
graph=FALSE)

## End(Not run)
```

VIII1 *CONTRAST OF INDEPENDENCE IN QUALITATIVE VARIABLES ASSOCIATION - POLYTOMOUS VARIABLES*

Description

It determines if several samples taken from the same population or different populations, are independent and in case of dependency, the degree of association in polytomous variables.

Usage

```
VIII1(data, variables, group=NULL, strata=NULL, test=c(1,2,3),
graph=TRUE, compress=TRUE, xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

`data` Data file.

`variables` Variable or variables to which the contrast of homogeneity is to be performed.

`group` Variables together with the data are grouped for calculations. In case of selecting NULL, no grouping would be done, and this would be calculated considering all data of the selected variables.

`strata` Variable defining the strata.

`test` TYPE OF CONTRAST OF HOMOGENEITY.

1. Chi-square test.
2. G-test without the Williams’ correction.

3. G-test with the Williams' correction.

4. Chi-square test with Yates correction (Only for 2x2 tables).

12. Measures of association.

13. Coefficient of uncertainty.

14. Cochran tests (without correction) (only tables 2x2xk).

15. Cochran-Mantel-Haenszel test (with correction) (only tables 2x2xk).

16. Calculating risk ratios per stratum (only 2x2xk.

17. Breslow & Day Test with Tarone correction. (Only 2x2xk).

`graph` GRAPH. If TRUE, the graph of the contingency table is shown. This chart does not allow accents and the categories are sorted by alphabetical order. The example of the function VIII1 shows how to sort the categories.

`compress` GRAPH. Logical value, if FALSE the space between rows and columns is chosen so that the total of heights and widths of the rows and columns are equal. If TRUE, the space between rows and columns is fixed and, therefore, the graph is more compressed.

`xlim` GRAPH. Vector with the X axis limits.

`ylim` GRAPH. Vector with the Y axis limits.

`spacing` GRAPH. Space between the bars, both horizontal and vertical, which can be extended by changing the value "sp=0" to a larger number. See the function spacings for more options.

`keep_aspect_ratio`
GRAPH. Logical value, if TRUE the height and width of the graph are the same.

`xscale` GRAPH. The categories bar width.

`yspace` GRAPH. Vertical spacing between bars that can be expressed in different units as "lines", "cm", "mm", "inches", etc. See funtion unit for more details.

`main` GRAPH. Main title of the graph.

`sub` GRAPH. Subtitle of the graph.

`residuals_type` GRAPH. It specifies the type of residuals but, for the moment, this only allows the Pearson residuals.

`shade` GRAPH. Logical value that if TRUE are the results of the statistical Chi square of Pearson and, in addition, the categories that are significantly different are shaded.

`gp_axis` GRAPH. It defines the line type, thickness, color, etc., of the line of the bars of the categories.

`grid.edit` GRAPH. Logical value that if TRUE allows to highlight a category with a different color.

`text.grid` GRAPH. Text that allows to identify the variables and categories of variables to highlight, for example "rect:Parents=Yes,Sex=Male". If text.grid is FALSE, it is not used. See one of the examples for more details.

`gp.grid` GRAPH. Object of the class gpar which allows to define the transparency and the color of the category to highlight.

labeling	GRAPH. It allows to set the labels for the categories on the left with "labeling_left", a framework with "labeling_cboxed", change from left to right and from top to bottom and put a framework with "labeling_lboxed", etc. See function labeling_border for more details.
labeling_args	GRAPH. It allows to specify the distance at which the category labels, legends, etc., are placed. See one of the examples for more details.
margins	GRAPH. Object of type unit with four components that delimit the margins from the top, right, down and left on the chart. See one of the examples for more details.
legend_width	GRAPH. Margin of the legend to the right edge.
gp_varnames	GRAPH. Size, format and font type letter of the legends of variables.
gp_labels	GRAPH. Size, format and font type letter of each of the categories of variables.
main_gp	GRAPH. Size, format, and font type letter of the main title.
sub_gp	GRAPH. Size, format, and font type letter of the subtitle.
legend	GRAPH. It allows to modify many aspects of the legend as the font size, font type, number of decimal places, change the text, etc.
gp	GRAPH. It allows to change the scale of color in the legend and also has different palettes: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" and "shading_binary".
file	TXT FILE. Output file name.

Details

VIII. CONTRAST OF INDEPENDENCE AND QUALITATIVE VARIABLES ASSOCIATION

One of the main purposes of the statistics is to determine if there is an association between two or more characters or variables of the same sample. In addition to know whether two or more variables are independent or dependent, it is also interesting to know what is the strength of this association.

In general, it should be said that the formula of approximation for the study of independence is the measurement of the differences between the observed frequencies and the frequencies that are expect to find if the variables were independent.

VIII.1. POLYTOMOUS VARIABLES

VIII.1.1. Tests of independence

VIII.1.1.1. Pearson's chi-squared test and Likelihood Ratio G-test

They are the most commonly used contrasts, although they tend not to be used for dichotomous variables, since there are specific tests for them.

Operationally the calculation of both statistical is equal to the contrast of homogeneity of samples (see paragraphs VII.1.1.1 and VII.1.1.2 in the section *details* of the function VII1), except that in this case, rows and columns relate to different variables of the same sample.

VIII.1.2. Measures of association

Measuring the strength of this association may be based on the calculation of distances between variables or in the amount of variability explained if the frequencies of a variable are calculated in function of the other variable.

VIII.1.2.1. V de Cramer

The coefficient *V* of Cramer is the following:

$$V = \sqrt{\frac{\chi^2}{n(min(f,c)-1)}}$$

being *f* and *c* the number of categories in rows and columns respectively and *n* the number of cases.

A value of 0 indicates total independence of the two variables, while a value of 1 indicates perfect association. It has the advantage to analyze contingency tables of any size.

VIII.1.2.2. Contingency Coefficient

It is also a measure associated to χ^2. The formula is:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}}$$

where *n* is the number of elements in the sample.

This coefficient takes values between 0 and 1. A value near zero indicates independence, and next to one indicates strong dependence. It can never take the limit value 1, even when there is total dependence.

VIII.1.2.3. Coefficient of Uncertainty

It is a measure of association based on the proportional error reduction, with a range between 0 and 1, from the measurement of a variable based on the knowledge of the other. That is, it indicates the percentage of the variability explained by the association of variables.

There are two different methods for performing this test:

1. Whereas an independent variable and one dependent. In this case two different statistical would be calculated as the dependent variable can consider both rows and columns. The measures of association are assuming the variable Y in columns and X in rows:

Dependent variable columns

$$I_{y|x} = \frac{(I_x + I_y - I_{xy})}{I_y}$$

Dependent variable rows

$$I_{x|y} = \frac{(I_x + I_y - I_{xy})}{I_x}$$

where

$$I_y = -\sum_{j=1}^{k} \frac{n_j}{n} ln \frac{n_j}{n}$$

$$I_x = -\sum_{i=1}^{h} \frac{n_i}{n} ln \frac{n_i}{n}$$

$$I_{xy} = -\sum_{i=1}^{h} \sum_{j=1}^{k} \frac{n_{ij}}{n} ln \frac{n_{ij}}{n}$$

being *h* and *k* the number of rows and columns, *n* the sample size, n_i and n_j the marginal frequencies of row and column, and n_{ij} joint frequencies.

2. Without considering dependence among variables

$$I_{ind} = 2\frac{(I_x + I_y - I_{xy})}{I_x + I_y}$$

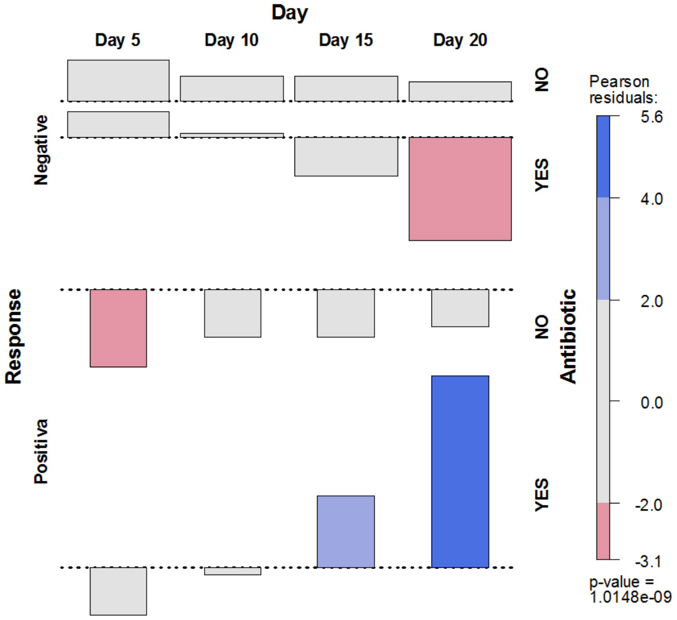
FUNCTIONS

The function assoc to make the graph (Meyer et al., 2006; 2013) was used. For more details on how to use this function, refer to the function help reference and/or Guisande & Vaamonde (2012).

EXAMPLE

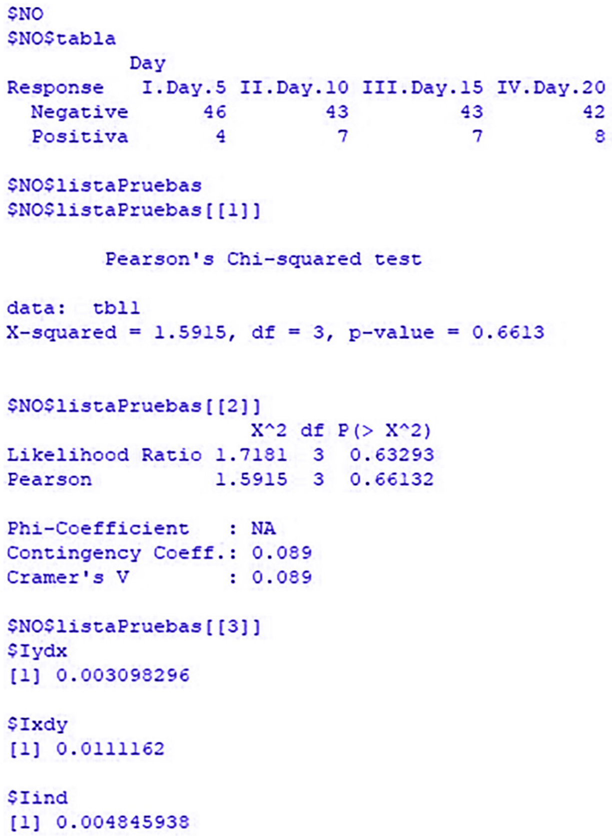
In the example of the function VII5 a study was conducted to determine if an antibiotic causes an allergic response in the skin.

Figure VIII1 Contingency table of the antibiotic and response variables, along the time.



The contrast of homogeneity showed that, both in the case of people treated with the antibiotic and those who were not provided, there were significant differences in the symptoms of allergy detected in the skin over the study period. It will be determined if there is indeed a relationship between time and the symptoms of allergy. The response variable is dichotomous to the antibiotic (positive or negative response), but the variable time is not dichotomized since there is data response at days 5, 10, 15 and 20 of the study.

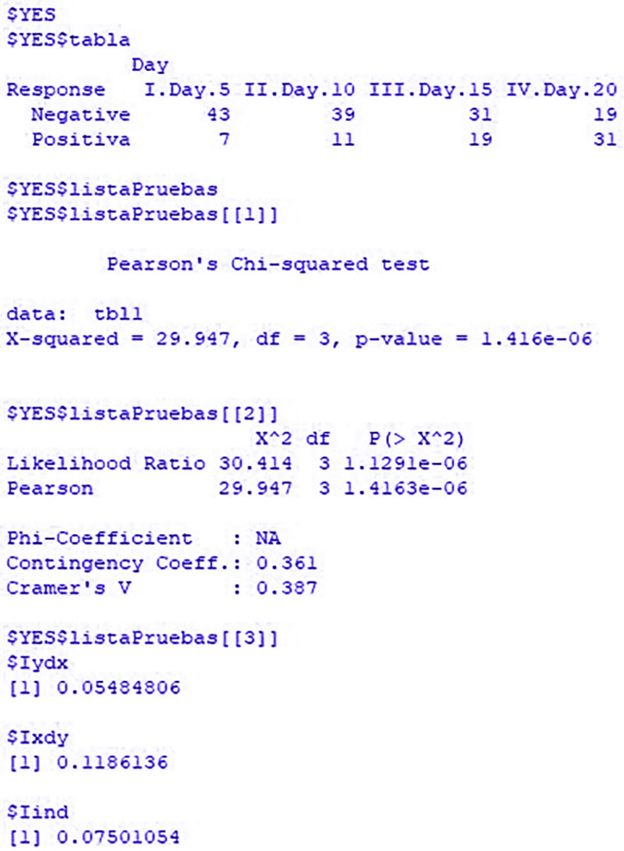
This chart does not allow accents and orderly categories in alphabetical order. The names of the columns are sorted with some sort of code in the input data file, to then leave these categories in the graph in the desired order. To change the labels, in order to include the accents, the argument *labeling_args* is used and then *set_varnames* is specified for the legends and *set_labels* for the labels of the categories, as shown in the script.



```
$NO
$NO$tabla
          Day
Response   I.Day.5 II.Day.10 III.Day.15 IV.Day.20
  Negative      46        43         43        42
  Positiva       4         7          7         8

$NO$listaPruebas
$NO$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 1.5915, df = 3, p-value = 0.6613


$NO$listaPruebas[[2]]
                    X^2 df P(> X^2)
Likelihood Ratio 1.7181  3  0.63293
Pearson          1.5915  3  0.66132

Phi-Coefficient   : NA
Contingency Coeff.: 0.089
Cramer's V        : 0.089

$NO$listaPruebas[[3]]
$Iydx
[1] 0.003098296

$Ixdy
[1] 0.0111162

$Iind
[1] 0.004845938
```

The joint analysis of the data shows that significant differences exist in the allergic response over time depending on whether an antibiotic is provided (Figure VIII1, p < 0.001). These differences were significant mainly in the day 20 where, to the people who were given the antibiotic, the positive responses were significantly higher and the negative responses were significantly lower (Figure VIII1).

The separate analysis shows that the group that was not given antibiotic, whose results are shown below, the null hypothesis of independence ($\chi^2_3 = 1.59$, p = 0.661) is accepted. Therefore, in spite of that the contrast of homogeneity showed significant differences in the allergic response observed in the skin over time, these observed differences are not sufficient to establish a significant dependency between time and the allergic response in those patients who were given a placebo instead of the antibiotic.



```
$YES
$YES$tabla
         Day
Response   I.Day.5 II.Day.10 III.Day.15 IV.Day.20
  Negative      43        39         31        19
  Positiva       7        11         19        31

$YES$listaPruebas
$YES$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 29.947, df = 3, p-value = 1.416e-06


$YES$listaPruebas[[2]]
                    X^2 df   P(> X^2)
Likelihood Ratio 30.414  3 1.1291e-06
Pearson          29.947  3 1.4163e-06

Phi-Coefficient   : NA
Contingency Coeff.: 0.361
Cramer's V        : 0.387

$YES$listaPruebas[[3]]
$Iydx
[1] 0.05484806

$Ixdy
[1] 0.1186136

$Iind
[1] 0.07501054
```

On the contrary, in the group of people who received antibiotic, the null hypothesis of independence (χ^2_3 = 29.9, $p < 0.001$) is rejected, that is to say, there is a significant relationship between time and the allergic response observed in the skin.

Among the association coefficients the *V* of Cramer, the coefficient of uncertainty and the contingency coefficient are considered. The *phi* is not considered because the table is not 2x2.

The coefficient *V* of Cramer is 0.39 and the contingency of 0.36, which indicates weak association. The coefficient of uncertainty y-dependent is 0.1186, that is to say, the variable X (time) reduced by 11.86% uncertainty on the prediction of the variable Y (allergy). Of course, it is considered that the variable allergy is time-dependent, and not vice versa.

With these results it is concluded that the variables are significantly related to value of 0.39 and, therefore, as allergy (variable Y) depends on the time (variable X), it can be concluded that the time explains a 12% of allergy observed in skin in people who were given the antibiotic.

Value

A TXT file is obtained with the results of the contrasts and the graph that shows the contingency table, if the option of displaying it is selected.

References

Meyer, D., Zeileis, A. & Hornik, K. (2006) The strucplot framework: Visualizing multiway contingency tables with vcd. *Journal of Statistical Software*, 17: 1-48.

Examples

```
## Not run:

data(ZVIII1)

#Allergic response of an antibiotic over time

VIII1(data=ZVIII1, variables=c("Response", "Day"), group="Antibiotic",
test=c(1,12,13), labeling_args =list(set_labels = list(Day= c("Day 5", "Day 10",
"Day 15", "Day 20"))), gp_varnames = gpar(fontsize=15,fontface=2),
gp_labels=gpar(fontsize = 13, fontface = 2))


## End(Not run)
```

VIII2 *CONTRAST OF INDEPENDENCE AND QUALITATIVE VARIABLES ASSOCIATION - DICHOTOMOUS VARIABLES*

Description

It determines if several samples taken from the same population or from different populations are independent and in case of dependency, the degree of association in dichotomous variables.

Usage

```
VIII2(data, variables, group=NULL, strata=NULL, test=c(1,2,3), graph=TRUE,
compress=TRUE, xlim=NULL, ylim=NULL, spacing=spacing_conditional(sp=0),
keep_aspect_ratio=FALSE, xscale=0.9, yspace=unit(0.5, "lines"), main=NULL,
sub=NULL, residuals_type="Pearson", shade=TRUE, gp_axis=gpar(lty = 3, lwd=2,
col="black"), grid.edit=FALSE, text.grid="rect:", gp.grid = gpar(fill="red"),
labeling=labeling_border,  labeling_args=list(), margins=unit(3, "lines"),
legend_width=unit(5, "lines"), gp_varnames=gpar(fontsize=12, fontface=1),
gp_labels=gpar(fontsize=12, fontface=1), main_gp=gpar(fontsize=20, fontface=2),
sub_gp=gpar(fontsize=15, fontface=1), legend = legend_resbased(fontsize=12,
fontfamily="Arial", x= unit(1, "lines"), y=unit(0.1,"npc"), height=unit(0.8, "npc"),
width=unit(0.7, "lines"), digits=2, check_overlap=TRUE, text=NULL, steps=200,
ticks=10, pvalue=TRUE), gp= shading_hcl, file="Output.txt")
```

Arguments

`data` Data file.

`variables` Variables whose association is to be determined.

`group` Variables together with the data are grouped to do the calculations. In the case of selecting NULL, there would be no grouping and this would be calculated considering all the data of the selected variables.

`strata` Variable defining the strata.

`test` TYPE OF CONTRAST OF HOMOGENEITY.

1. Chi-square test.

2. G-test without the Williams' correction.

3. G-test with the Williams' correction.

4. Chi-square test with Yates correction (Only for 2x2 tables).

12. Measures of association.

13. Coefficient of uncertainty.

14. Cochran tests (without correction) (only tables 2x2xk).

15. Cochran-Mantel-Haenszel test (with correction) (only tables 2x2xk).

16. Calculating risk ratios per stratum (only 2x2xk.

17. Breslow & Day Test with Tarone correction. (Only 2x2xk).

`graph` GRAPH. If TRUE, the graph of the contingency table is shown. This chart does not allow accents and the categories sorted by alphabetical order. In the example, the function VIII1 shows how to sort the categories and include the accents.

`compress` GRAPH. Logical value, if FALSE the space between rows and columns is chosen so that the total of heights and widths of the rows and columns are equal. If TRUE, the space between rows and columns is fixed and, therefore, the graph is more compressed.

`xlim` GRAPH. Vector with the X axis limits.

`ylim` GRAPH. Vector with the Y axis limits.

`spacing` GRAPH. Space between the bars, both horizontal and vertical, which can be extended by changing the value "sp=0" to a larger number. See the function spacings for more options.

`keep_aspect_ratio`
GRAPH. Logical value, if TRUE the height and width of the graph are the same.

`xscale` GRAPH. The categories bar width.

`yspace` GRAPH. Vertical spacing between bars that can be expressed in different units as "lines", "cm", "mm", "inches", etc. See funtion unit for more details.

`main` GRAPH. Main title of the graph.

`sub` GRAPH. Subtitle of the graph.

`residuals_type` GRAPH. It specifies the type of residuals but, for the moment, this only allows the Pearson residuals.

`shade` GRAPH. Logical value that if TRUE are the results of the statistical Chi square of Pearson and, in addition, the categories that are significantly different are shaded.

gp_axis GRAPH. It defines the line type, thickness, color, etc., of the line of the bars of the categories.

grid.edit GRAPH. Logical value that if TRUE allows to highlight a category with a different color.

text.grid GRAPH. Text that allows to identify the variables and categories of variables to highlight, for example "rect:Parents=Yes,Sex=Male". If text.grid is FALSE, it is not used. See one of the examples for more details.

gp.grid GRAPH. Object of the class gpar which allows to define the transparency and the color of the category to highlight.

labeling GRAPH. It allows to set the labels for the categories on the left with "labeling_left", a framework with "labeling_cboxed", change from left to right and from top to bottom and put a framework with "labeling_lboxed", etc. See function labeling_border for more details.

labeling_args GRAPH. It allows to specify the distance at which the category labels, legends, etc., are placed. See one of the examples for more details.

margins GRAPH. Object of type unit with four components that delimit the margins from the top, right, down and left on the chart. See one of the examples for more details.

legend_width GRAPH. Margin of the legend to the right edge.

gp_varnames GRAPH. Size, format and font type letter of the legends of variables.

gp_labels GRAPH. Size, format and font type letter of each of the categories of variables.

main_gp GRAPH. Size, format, and font type letter of the main title.

sub_gp GRAPH. Size, format, and font type letter of the subtitle.

legend GRAPH. It allows to modify many aspects of the legend as the font size, font type, number of decimal places, change the text, etc.

gp GRAPH. It allows to change the scale of color in the legend and also has different palettes: "shading_hcl", "shading_hsv", "shading_Friendly", "shading_max", "shading_sieve" and "shading_binary".

file TXT FILE. Output file name.

Details

VIII. CONTRAST OF INDEPENDENCE AND QUALITATIVE VARIABLES ASSOCIATION

VIII.2. DICHOTOMOUS VARIABLE

VIII.2.1. Tests of independence

VIII.2.1.1. Yates'correction and Fisher's test

To contrast independence in dichotomous variables, when there is only one table 2x2, the test χ^2 with Yates' correction can be used, which has already been explained in paragraph VII.1.2.1 of the function VII2.

Fisher's exact test, which is briefly explained in the paragraph VII.1.2.2 function VII2, is also used when the contingency table is 2x2 and should be used instead of the test χ^2 when the contingency table contains any cell with expected frequencies under 5.

VIII.2.1.2. The Cochran-Mantel-Haenszel test

As with the Fisher's test, the evidence of Cochran (1954) and Mantel-Haenszel (1959) also serve to contrast the independence in a 2x2 Contingency Table.

However, they have the advantage over Fisher's test that can also be used when there are several 2x2 tables. These contingency tables are of the type *K*x2x2, that is to say, there are so many 2x2 tables such as *K* strata.

In these tables, the aim is to study whether or not there is association between two dichotomous variables, when there is information on several strata. For example, to determine if the presence of cervical cancer is related to the presence of human papilloma virus in women of different ages or on the contrary are independent.

Therefore, multiple 2x2 tables (presence or non-presence of cancer and presence or non-presence of the virus, both dichotomous variables) would be obtained in several layers that would be the different classes of age at which the group of women studied was divided. In this example, analyze each 2x2 table separately with Fisher's test would not give information about the effect of age. The statistic of Cochran is as follows:

$$\chi^2_{Cochran} = \frac{\left(\sum_k n_k - \sum_k m_k\right)^2}{\sum_k \sigma^2_{n_k}}$$

where:

k = each of the strata

n_k = frequency observed in any of the boxes of the stratum *k* (only one and is always the same in all strata)

m_k = expected frequency corresponding ton_k

$$\sigma^2_{n_k} = \frac{n_{f1k} n_{f2k} n_{c1k} n_{c2k}}{n^3}$$

n_{f1k}, n_{f2k}, n_{c1k} y n_{c2k} they are the four marginal frequencies associated with the 2 X 2 tables of each stratum *k*.

The Mantel-Haenszel statistical is very similar to the Cochran, except for two things:

1) It uses the correction for continuity (midpoint is subtracted from the numerator before squaring).

2) It slightly changes the denominator of the variance, using $n^2(n-1)$ instead of n^3.

In both statistics, the Cochran's and Mantel-Haenszel, if the value of p is greater than 0.05, the hypothesis of independence is accepted and it is concluded that, after taking into account the effect of the strata, the two dichotomous variables are not associated.

VIII.2.2. Measures of association

In the case of 2x2 contingency tables, any of the three factors discussed above in section VIII.1.2 of the VIII1 function can be used: *V* Cramer, Contingency Coefficient and Coefficient of Uncertainty. Besides these three coefficients, the coefficient Phi (ϕ) can be used.

VIII.2.2.1. Phi

The range of this ratio is only between 0 and 1 in 2x2 contingency tables and therefore should only be used when the variables to compare are both dichotomous.

Phi (ϕ) is a measure based on the statistical χ^2. The formula is:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

where *n* is the number of elements in the sample.

VIII.2.2.2. Relative risk, odds ratio and Breslow-Day test and Tarone

When there are contingency tables *K*x2x2, mentioned above, the importance of association will be better estimated using the relative risk and odds ratio. These measures can also be used in tables 2x2, as well as the coefficients described above (*V* of Cramer, contingency coefficient, Coefficient of uncertainty and Phi), but in the case of the tables *K*x2x2 these measures have the advantage that they can be compared between the layers or strata.

Relative risk indicates how much more likely is the success of a first group against the second. In a 2 x 2 table as shown in the table, the risk or probability of the event considered in Group A and Group B would be:

	Group A	**Group B**	**Total**
Event	a	c	a+c
No event	b	d	b+d
Total	a+b	c+d	*n*

$$Risk(A) = \frac{a}{(a+b)} \quad Risk(B) = \frac{c}{(c+d)}$$

the ratio of these two quantities is the relative risk of the event (RR_{event}) and likewise it could be calculated the relative risk of the event or otherwise "no event" ($RR_{no-event}$)

$$RR_{event} = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} \quad RR_{no-event} = \frac{\frac{b}{(a+b)}}{\frac{d}{(c+d)}}$$

The ratio of both relative risks is equal to the odds ratio or Reason of predominances (*OR*, *Odds Ratio*):

$$OR = \frac{RR_{event}}{RR_{no-event}} = \frac{(a*d)}{(b*c)}$$

The interpretation of the relative risk is more intuitive than the reason of predominances. However the odds ratio is widely used, mainly because of its connection with the Logistic Regression, which will be reflected in another chapter, with important applications in epidemiology and other areas.

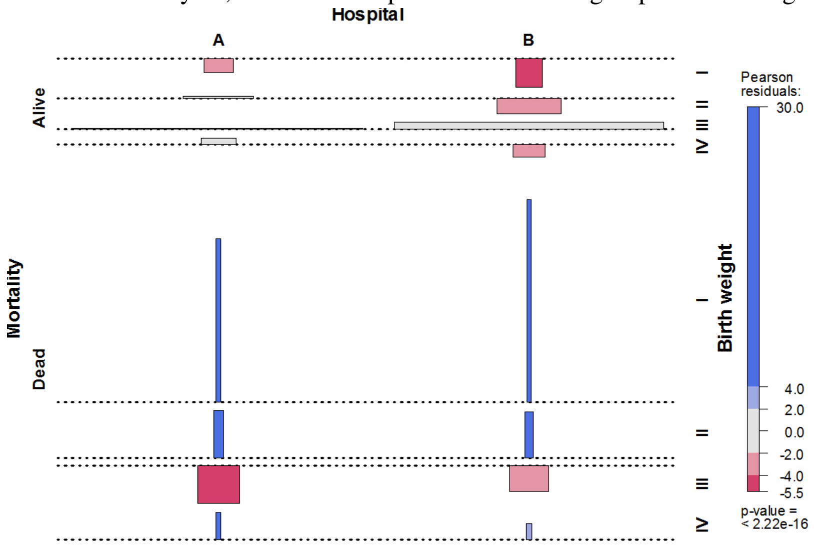
Breslow & Day test (1980) and the modification of this test of Tarone (1985) are homogeneity tests that compare odds ratios between different strata, in *K*x2x2 contingency tables. With this information it is possible to determine whether the degree of association between two dichotomous variables varies according to the different strata. In the example mentioned above, this would inform about if the possible association between human papillomavirus and cervical cancer increases or becomes less intense with the age of the women.

EXAMPLE

The number of children who died after a year regarding his birth weight was analyzed in two different hospitals. The group I were newborns weighing less than 1.5 kg, group II between 1.5

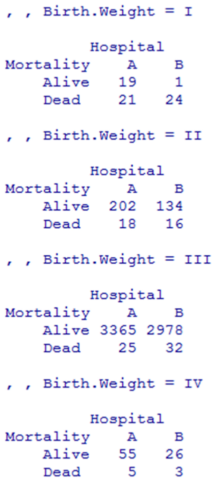
and 2.5 kg, group III between 2.5 and 4.2 kg and group IV over 4.2 kg. The objective was to determine whether there were differences in mortality between hospitals and tried to eliminate the possible effect of any birthweight infant mortality.

Figure VIII2 Graphical representation of the contingency table in the proportion of live babies and dead after a year, in the two hospitals and for each group of birth weight.



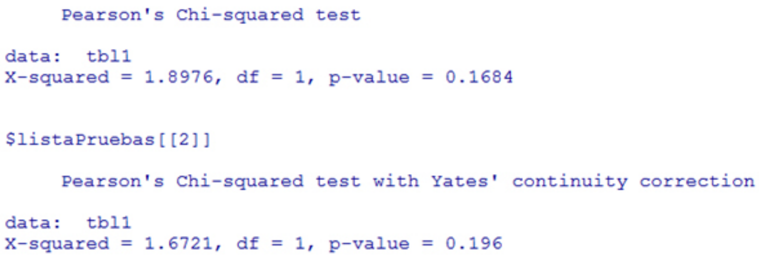
The joint analysis of the data shows that there are significant differences in mortality of babies (Figure VIII2, p < 0.001). In both hospitals the percentage of dead babies after one year is significantly higher in groups I and II (babies with less weight at birth, blue bars above the dotted line) and also the group IV (babies with a weight much higher than normal). On the contrary, the percentage of dead babies was significantly lower (bars below the dotted line) in the group III.

The first thing that one can observe when a TXT file with the results is opened are a series of double entry tables with summarized data.



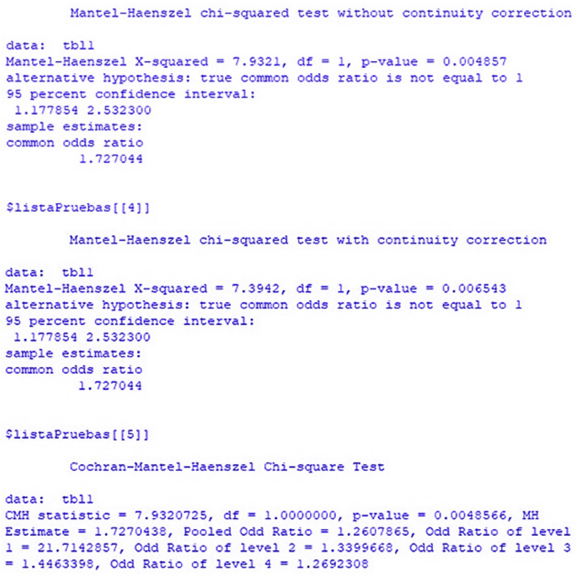
```
, , Birth.Weight = I

         Hospital
Mortality    A    B
    Alive   19    1
    Dead    21   24

, , Birth.Weight = II

         Hospital
Mortality    A    B
    Alive  202  134
    Dead    18   16

, , Birth.Weight = III

         Hospital
Mortality    A    B
    Alive 3365 2978
    Dead    25   32

, , Birth.Weight = IV

         Hospital
Mortality    A    B
    Alive   55   26
    Dead     5    3
```

Then a series of different analyses results can be observed. The χ^2 without (χ^2_1 = 1.9, p = 0.168) and with Yates' correction (χ^2_1 = 1.6, p = 0.196) are not significant, so it was concluded that there is no association between the two variables, which means that without considering the birthweight - there is no difference in the percentage of live and dead babies between hospitals.



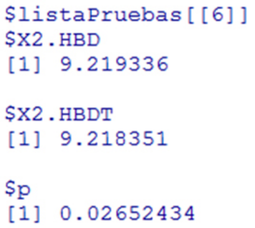
```
     Pearson's Chi-squared test

data:  tbl1
X-squared = 1.8976, df = 1, p-value = 0.1684


$listaPruebas[[2]]

     Pearson's Chi-squared test with Yates' continuity correction

data:  tbl1
X-squared = 1.6721, df = 1, p-value = 0.196
```

Cochran tests (without continuity correction, χ^2_1 = 7.9 , p = 0.004) and Mantel-Haenszel (with continuity correction, χ^2_1 = 7.39 , p = 0.006) were significant, indicating that if it is considered the birth weight, mortality depends on the hospital, it ceases to be independent. They also show the values of the odds ratio, that explain this relationship: in level 1 of the stratification factor (babies with birth weight < 1.5 kg) predominance of deaths on not dead at hospital A is 0.046 times the predominance at hospital B (22 times smaller); in the remaining levels of weight the differences are much smaller.



```
        Mantel-Haenszel chi-squared test without continuity correction

data:  tbl1
Mantel-Haenszel X-squared = 7.9321, df = 1, p-value = 0.004857
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.177854 2.532300
sample estimates:
common odds ratio
         1.727044


$listaPruebas[[4]]

        Mantel-Haenszel chi-squared test with continuity correction

data:  tbl1
Mantel-Haenszel X-squared = 7.3942, df = 1, p-value = 0.006543
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.177854 2.532300
sample estimates:
common odds ratio
         1.727044


$listaPruebas[[5]]

        Cochran-Mantel-Haenszel Chi-square Test

data:  tbl1
CMH statistic = 7.9320725, df = 1.0000000, p-value = 0.0048566, MH
Estimate = 1.7270438, Pooled Odd Ratio = 1.2607865, Odd Ratio of level
1 = 21.7142857, Odd Ratio of level 2 = 1.3399668, Odd Ratio of level 3
= 1.4463398, Odd Ratio of level 4 = 1.2692308
```

Breslow's test (HBD = 9.2) and the correction of Tarone (HBDT = 9.2) are used to check if there are significant differences between strata. In this case, the p-value associated with these statistics is 0.026, which reject the null hypothesis that there are no differences between strata. This means that birth weight influences the mortality of babies during their first year of life.



```
$listaPruebas[[6]]
$X2.HBD
[1] 9.219336

$X2.HBDT
[1] 9.218351

$p
[1] 0.02652434
```

Value

A TXT file is obtained with the results of the contrasts and the graph that shows the contingency table, if the option of displaying it is selected.

References

Breslow, N.E. & Day, N.E. (1980) *Statistical methods in cancer research*. Volume 1. The analysis or case-control studies. Lyon: International Agency for Research on Cancer.

Cochran, W.G. (1954) Some methods for strengthening the common χ^2 tests. *Biometrics*, 10: 417-451.

Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22: 719-748.

Meyer, D., Zeileis, A. & Hornik, K. (2006) The strucplot framework: Visualizing multiway contingency tables with vcd. *Journal of Statistical Software*, 17: 1-48.

Tarone, R.E. (1985) On heterogeneity tests based on efficient scores. *Biometrika*, 72: 91-95.

Examples

```
## Not run:

data(ZVIII2)

#Mortality at birth

VIII2(data=ZVIII2, variables=c("Mortality","Hospital"), strata="Birth.Weight",
test=c(1,4,14,15,16,17), gp_labels=gpar(fontsize = 14, fontface = 2),
gp_varnames=gpar(fontsize = 16, fontface = 2),
labeling_args =list(set_varnames=c(Birth.Weight="Birth weight" )))


## End(Not run)
```

X1 *SIMPLE BIVARIATE CORRELATIONS*

Description

An analysis of correlation between two variables is applied using the coefficients of Pearson, Spearman and Kendall. In addition, the relationship between the dependent variable and the independent can be plotted, and also be distinguished between groups.

Usage

```
X1(data, varX,  varY, test=c("Pearson","Spearman", "Kendall"),
group=NULL, ResetPAR=TRUE, PAR=NULL, YLAB=NULL, XLAB=NULL,
CEXPCH=1.3, COLOR=NULL, PCH=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL,
TEXT=NULL, file="Output.txt")
```

Arguments

`data` Data file.

`varX` Independent variable.

`varY` Dependent variable.

`test` One, two or three correlation coefficients may be chosen: Pearson, Spearman and/or Kendall.

`group` Variable in the case of wanting to estimate the correlations by groups.

`ResetPAR` If FALSE, the default conditions of the function PAR are not placed and are those defined by the user in previous graphs.

`PAR` It accesses the function PAR which allows to modify many different aspects of the graph.

`YLAB` Legend of the Y axis.

`XLAB` Legend of the X axis.

`CEXPCH` Size of the symbols on the chart.

`COLOR` It allows to modify the colors of the symbols, but in the event that *group* is not NULL, should be as many different groups as the *group* variable has.

`PCH` Vector with graphic symbols. If NULL, these are automatically calculated starting with the symbol 15.

`LEGEND` It allows to add a legend to the graph.

`AXIS` It allows to add axes to the graph.

`MTEXT` It allows to add text in the margins of the graph.

`TEXT` It allows to add text in any area of the inner part graphic.

`file` TXT FILE. Output file name.

Details

X. CONTRASTS OF INDEPENDENCE AND ASSOCIATION IN QUANTITATIVE VARIABLES

One of the most frequent objectives when analizing data is to study the degree of dependence or association between variables. The correlation is the overarching theoretical framework within which we can study these aspects.

It is said that a variable *y* depends on another variable *x* when the value of *x* determines to some extent the value of *y*. It may be the case that the value of *y* is completely determined by the value of *x*. Then, if the value of the variable *x* is known, there is a functional relationship that defines the perfect form of value that *y* would take. In this case it seems that between the two variables there is a relationship of functional dependence. In this case it would seem that there is a relationship between the two variables of functional dependence. But in practice, most often will be the situations in which the value of the variable and it is not perfectly defined by the value of *x*, but there is a certain degree of relationship between the two that allows an approximation to the value of *y*. In this case, a relation of dependence between the two random variables.

A series of statistics that are called correlation coefficients are used to analyze the degree of dependence or association between variables. The correlation coefficients indicate if the variables are associated or not (if independent) and also, in what sense that association or correlation between variables (positive or negative) is given. They also report on the intensity of the correlation. Through various statistics, it is also possible to know whether the correlation between variables is significant or not.

X.1. SIMPLE BIVARIATE CORRELATIONS

X.1.1. Pearson's correlation coefficient (*r*)

The Pearson correlation coefficient (*r*) indicates the strength of the linear relationship between the two variables. It is necessary to assume, therefore, that any dependency relationship that exists between the two variables is linear. That only serves to quantitative variables.

The coefficient *r* of Pearson is calculated according to the following formula, which is based on the covariance between the two variables:

$$r_{y_1,y_2} = \frac{\sum_{i=1}^{n} [(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2))]}{\sqrt{\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)^2 \sum_{i=1}^{n}(y_{i2} - \bar{y}_2)^2}}$$

or what is the same

$$\frac{S_{xy}}{S_x S_y}$$

that is, the covariance divided by the product of the standard deviations.

The values of this ratio are spread in a range between -1 and 1. The closer to -1 or 1, the stronger is the correlation between the variables. Values close to 0 indicate the absence of correlation and, therefore, of dependence.

For the significance test of the Pearson's coefficient, it is usually set as H_0 that *R* = 0, using the correlation coefficient *r* obtained from the sample as estimator of the true correlation coefficient population *R*. In addition, it is necessary that the joint probability distribution of both variables is Normal.

$$t = \frac{r}{s_r}$$

where s_r is the standard error of the estimator *r*:

$$s_r = \sqrt{\frac{(1 - r^2)}{(n - 2)}}$$

X.1.2. Spearman’s correlation coefficient (ρ)

Among the non-parametric methods, the Spearman coefficient (ρ) is the most commonly used. As in all the non-parametric methods, it is not necessary to make any assumptions about the distribution that follows the data. This is a statistic that is calculated in a similar way to the Pearson’s coefficient, but allocating ranges to the data. When there are several data with the same value, the average of the corresponding ranges is assigned to all of them.

This ratio works for any type of association between variables that follow a monotonic function type. Obviously, this is due to the fact that an ordinal scale is used, so this property is useful for any other coefficient that uses ranges to be calculated. This is an important difference compared to the Pearson’s coefficient, which only responds well if the relationship between two variables is linear. In addition, it serves both quantitative as qualitative ordinal variables.

It is calculated by using the following formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference, for each data or observation, between the ranges that take the values of the two variables.

The range of values that takes this coefficient ranges from - 1 to 1. The closer of these ends, the stronger is the degree of association, and the sign indicates whether the functional relationship between two variables is ascending or descending, in the same way as the Pearson’s coefficient.

To determine the significance of this coefficient the following statistic, whose distribution approaches the *t* Student, is used and when the sample size is greater than 20 units:

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

In the same way as Pearson’s coefficient, it is usually set as H_0 that ρ = 0.

X.1.3. Coefficient tau of Kendall

This coefficient is particularly suitable for variables expressed in ordinal form. Therefore it can also be used after assigning ranges to the values of the variables.

There is a formula for its calculation that does not take into account "ties" between ranges of data, it is denoted by τ_A, and is calculated as follows:

$$\tau_A = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

where n_c is the number of concordant pairs of ranges and n_d is the number of discordant pairs of ranges.

In the event that there is data in which there are "ties" between ranges, this new formula would be used:

$$\tau_B = \frac{n_p - n_q}{\sqrt{(n_p + n_q + n_{E(X)})(n_p + n_q + n_{E(Y)})}}$$

where *p* indicates concordance or not investment, *q* discordance or investment, and *E* tie considering all pairs of cases.

The values of this coefficient are between -1 and 1, and its properties and interpretation are similar to the coefficients of Pearson and Spearman.

The significance is contrasted by means of statistical that are distributed according to a Normal. In the case of τ_A the following formula is applied:

$$z = \frac{3(n_c - n_d)}{\sqrt{\frac{n(n-1)(2n+5)}{2}}}$$

In the case of τ_B the following formula would be used:

$$z = \frac{n_c - n_d}{\sqrt{\upsilon}}$$

where:

$$\upsilon = \frac{\upsilon_0 - \upsilon_t - \upsilon_u}{18 + \upsilon_1 + \upsilon_2}$$

which in turn:

$$\upsilon_0 = n(n-1)(2n+5)$$

$$\upsilon_t = \sum t_i(t_i - 1)(2t_i + 5)$$

$$\upsilon_u = \sum\nolimits_j u_j(u_j - 1)(2u_j + 5)$$

$$\upsilon_1 = \sum\nolimits_i t_i(t_i - 1) \sum\nolimits_j \frac{u_j(u_j - 1)}{2n(n-1)}$$

$$\upsilon_2 = \sum\nolimits_i t_i(t_i - 1)(t_i - 2) \sum\nolimits_j \frac{u_j(u_j - 1)(u_j - 2)}{9n(n-1)(n-2)}$$

The H_0 for the significance of this coefficient is that its value is 0 and therefore no association between variables.

X.1.4. Gamma coefficient γ

The most appropriate coefficient when the data present a high overlap of ranges. Again, the range of values that this coefficient can take is between -1 and 1, with -1 perfect negative association, 1 perfect positive association and 0 no association. The formula is as follows:

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

where n_c is the number of concordant pairs of ranges and n_d is the number of discordant pairs.

The significance is checked by using the following statistical value which is distributed by following a Normal

$$z = \gamma \sqrt{\frac{n_c - d_d}{N(1 - \gamma^2)}}$$

As with the rest of coefficients, the H_0 is that the value of the coefficient is equal to 0, and therefore there is no association between the variables.

FUNCTIONS

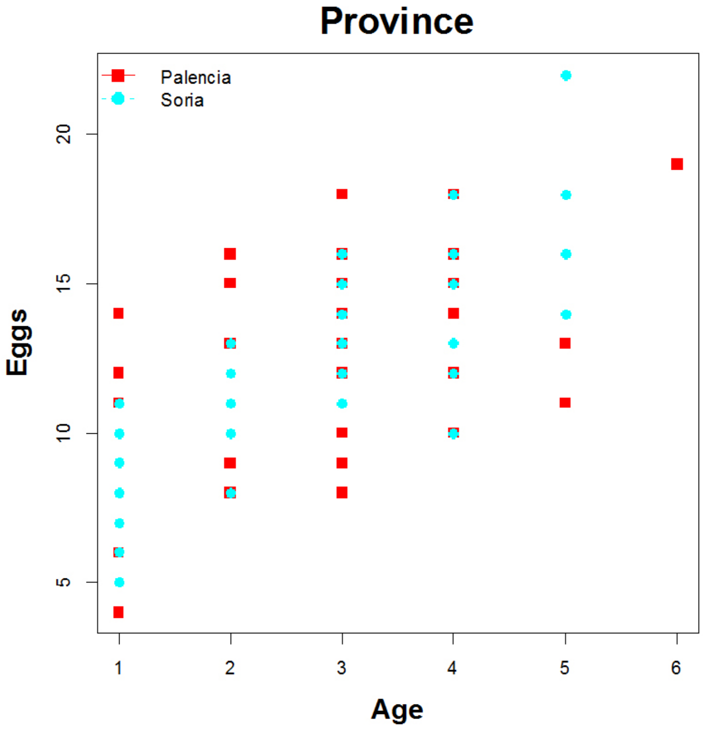
To estimate the correlations, the function used is as follows rcor.test (Rizopoulos 2006; 2013).

EXAMPLE

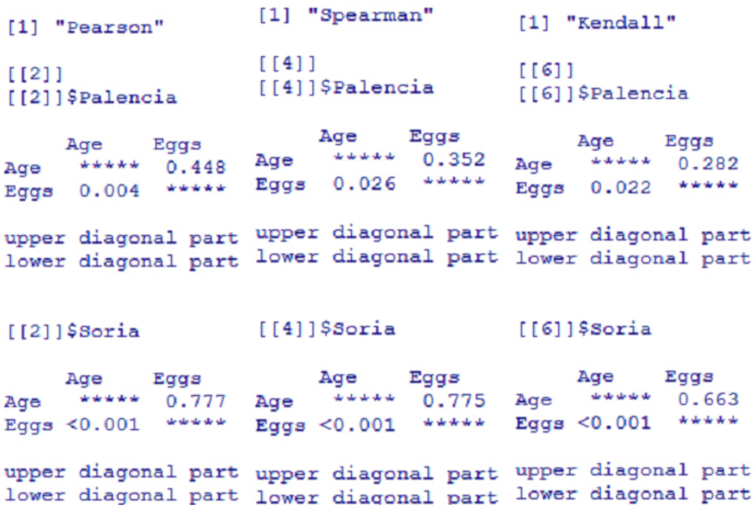
Data from a sampling in nests of a small bird species, the common quail. In this sample, the number of eggs per female, posts during a breeding season is counted in 50 females marked whose age was known. At the same time the average temperature during the same breeding season and the availability of food for females (whose values are percentages are given for the maximum measured) were recorded. The same sampling was repeated in the next breeding season with a different sample of females in the same population. The objective of the research is to discover if there is a relationship between the age of the female and the number of eggs per clutch. Estimates are made for each of the Provinces sampled.

Figure X.1 shows that there is a clear positive relationship between the age of the female and the number of eggs per clutch in the two provinces.

Figure X.1 Relationship between the age of the female quail and the number of eggs per clutch, for each of the provinces.



The results table shows each of the correlation coefficients in the upper-right half of the table, and in the lower half left the probability. It is noted that, in all cases, the assumption of independence is rejected, because the probability is always less than 0.05. Therefore, in both provinces and considering the three coefficients, it is concluded that there is a relationship of dependency between the age of the female and the number of eggs. This association seems to be stronger in the province of Soria where, for example, Pearson's correlation coefficient is 0.78 , while in Palencia is 0.45. The coefficients of Spearman and Kendall also show this greater association between both variables in one province than in the other.



```
[1] "Pearson"             [1] "Spearman"             [1] "Kendall"

[[2]]                     [[4]]                      [[6]]
[[2]]$Palencia            [[4]]$Palencia             [[6]]$Palencia

        Age    Eggs               Age    Eggs                Age    Eggs
Age     *****  0.448      Age     *****  0.352       Age     *****  0.282
Eggs    0.004  *****      Eggs    0.026  *****       Eggs    0.022  *****

upper diagonal part       upper diagonal part        upper diagonal part
lower diagonal part       lower diagonal part        lower diagonal part


[[2]]$Soria               [[4]]$Soria                [[6]]$Soria

        Age    Eggs               Age    Eggs                Age    Eggs
Age     *****  0.777      Age     *****  0.775       Age     *****  0.663
Eggs   <0.001  *****      Eggs   <0.001  *****       Eggs   <0.001  *****

upper diagonal part       upper diagonal part        upper diagonal part
lower diagonal part       lower diagonal part        lower diagonal part
```

Value

A TXT file is obtained with the results of the correlations and can represent a graph relating the dependent variable with the independent, differentiating between the groups of the grouping variable.

References

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, **17**, 1-25.

Rizopoulos, D. (2013) Latent Trait Models under IRT. R package version 1.0-0. Available at: http://CRAN.R-project.org/package=ltm.

Examples

```
## Not run:

data(ZX1)

X1(data=ZX1,varX="Age", varY="Eggs", group="Province")


## End(Not run)
```

X2 *PARTIAL CORRELATIONS*

Description

A partial correlation analysis is applied between two variables using the coefficients of Pearson and Kendall, and taking into account the effect of another variable. In addition the relationship between the dependent variable and the independent can be plotted, and also distinguished between groups.

Usage

```
X2(data, varX, varY, varZ, test=c("Pearson","Kendall"),
group=NULL,
ResetPAR=TRUE, PAR=NULL, YLAB=NULL, XLAB=NULL, CEXPCH=1.3, COLOR=NULL,
PCH=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL, file="Output.txt")
```

Arguments

`data`	Data file.
`varX`	Independent variable.
`varY`	Dependent variable.
`varZ`	Variable that the effects on the correlation wants to be tested.
`test`	One, or two correlation coefficients may be chosen: Pearson and/or Kendall.
`group`	Variable in the case of wanting to estimate the correlations by groups.
`ResetPAR`	If FALSE, the default conditions of the function PAR are not placed and are those defined by the user in previous graphs.
`PAR`	It accesses the function PAR which allows to modify many different aspects of the graph.
`YLAB`	Legend of the Y axis.
`XLAB`	Legend of the X axis.
`CEXPCH`	Size of the symbols on the chart.
`COLOR`	It allows to modify the colors of the symbols, but in the event that *group* is not NULL, should be as many different groups as the *group* variable has.
`PCH`	Vector with graphic symbols. If NULL, these are automatically calculated starting with the symbol 15.
`LEGEND`	It allows to add a legend to the graph.
`AXIS`	It allows to add axes to the graph.
`MTEXT`	It allows to add text in the margins of the graph.
`TEXT`	It allows to add text in any area of the inner part graphic.
`file`	TXT FILE. Output file name.

Details

X. CONTRASTS OF INDEPENDENCE AND ASSOCIATION IN QUANTITATIVE VARIABLES

X.2. PARTIAL CORRELATIONS

X.2.1. Pearson correlation coefficient

Sometimes a bivariate linear correlation can be masked by the effect of other variables beyond our control. For example, there may be a strong relationship between the size of the clutch of a certain bird species and the age of the female and, however, this correlation is masked by the influence of other variables in our data such as the average temperature during the breeding season. To study a simple correlation between two variables of interest, at the same time that the effect of other variables is controlled, the partial correlations are used.

If there are two variables *X* and *Y* on which the partial correlation is calculated controlling *k* variables *Z*, the method is based on calculating the Pearson correlation coefficient between two derived variables d_x and d_y such that:

d_x = residues of the linear regression of *x* with z_1, ..., z_k
d_y = residues of the linear regression of *y* withz_1, ..., z_k

It is possible to control, at the same time, the effect of one or more variables *Z*. If only one is controlled, it would be partial correlation of first order. If two are controlled, it would be partial correlation of second order, and so on.

The calculation of the partial correlations of first order is by using the following formula:

$$r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

where *r* is the Pearson correlation coefficient between the indicated variables.

As an example of greater order, the formula of the coefficient is shown in the case that they are controlling two variables (z_i and z_j) whereby is a partial correlation of second order. The higher-order is obtained following the same logic:

$$r_{xy,z_iz_j} = \frac{r_{xy,z_i} - r_{xz_j,z_i}r_{yz_j,z_i}}{\sqrt{(1 - r_{xz_j,z_i}^2)(1 - r_{yz_j,z_i}^2)}}$$

The test of significance of this coefficient, with the null hypothesis that the population value is zero, is performed by calculating the statistical *t* by means of the following formula:

$$T = r_{xy,k}\frac{\sqrt{m - k - 2}}{\sqrt{(1 - r_{xy,k}^2)}}$$

where *m* it is the number of cases (if it is not equal for all pairs of variables, the less of them must be taken) and *k* is the number of controlled variables that has been used.

X.2.2. Partial Kendall correlation coefficient ($T_{xy,z}$)

The basis of the partial correlations is the same as in the case of parametric statistics but a different statistic is needed which allows working without making any assumption about the distribution that our data follow. The main statistical that performs this function is the T of Kendall, which is essentially the same as the τ described in paragraph X.1.3 of the critical role X1, but adapted to the calculation of partial correlations.

The formula to compute it is as follows:

$$T_{xy,z} = \frac{T_{xy} - T_{xz}T_{yz}}{\sqrt{(1 - T^2_{xz})(1 - T^2_{yz})}}$$

where each bivariate correlation T is calculated in the same way that the τ as described in section X.1.3 function X1.

For the calculation of its significance, the following statistical which follows a Normal distribution is used:

$$z = \frac{3T_{xy,z}\sqrt{N(N-1)}}{\sqrt{2(2N+5)}}$$

The H_0 set, as usual for these statistics, is that the value of the population coefficient is 0.

FUNCTIONS

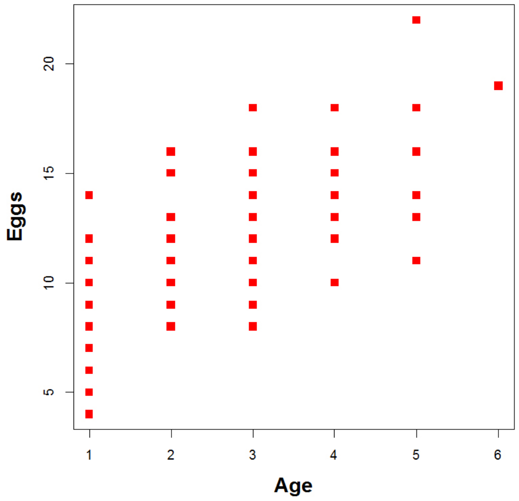
To estimate the correlations the following function rcor.test is used (Rizopoulos 2006; 2013).

EXAMPLE

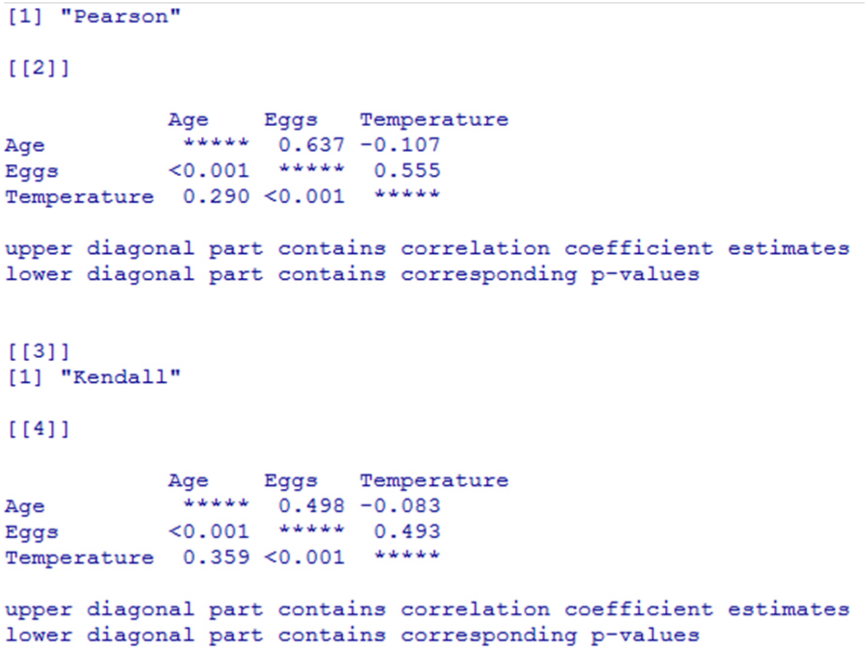
Data from a sampling in nests of a small bird species, the common quail. In this sample the number of eggs laid per female during the breeding season is counted in 50 tagged females whose age was known. At the same time the average temperature during the same breeding season and the availability of food for females (whose values are percentages are given for the maximum measured) were recorded. The same sampling was repeated in the next breeding season with a different sample of females in the same population. The objective of the research is to determine if temperature affects the possible relationship between the age of the female and the number of eggs per clutch. The calculations are performed without differentiating between provinces.

Figure X.2 shows that there is a clear positive relationship between female age and the number of eggs per clutch, considering the set of all provinces.

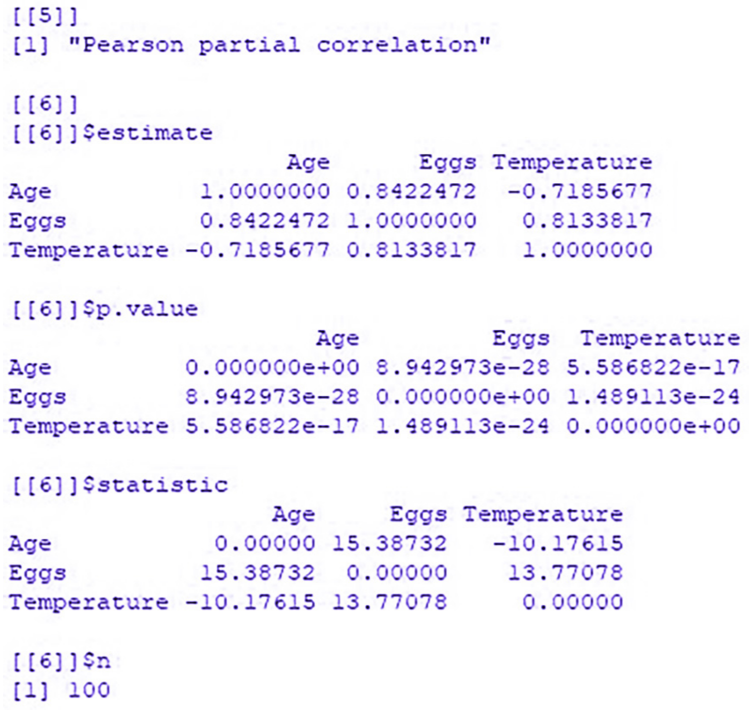
Figure X.2 Relationship between the age of the female quail and the number of eggs per clutch, considering all provinces.



In the Pearson coefficient, the upper right matrix shows that there is a correlation between the two variables of 0.637, and the lower left is the probability matrix. It is observed that the hypothesis of independence is rejected, since p < 0.001. Therefore, there is a significant correlation between female age and the number of eggs per female. The Kendall coefficient is 0.498 and p < 0.001.

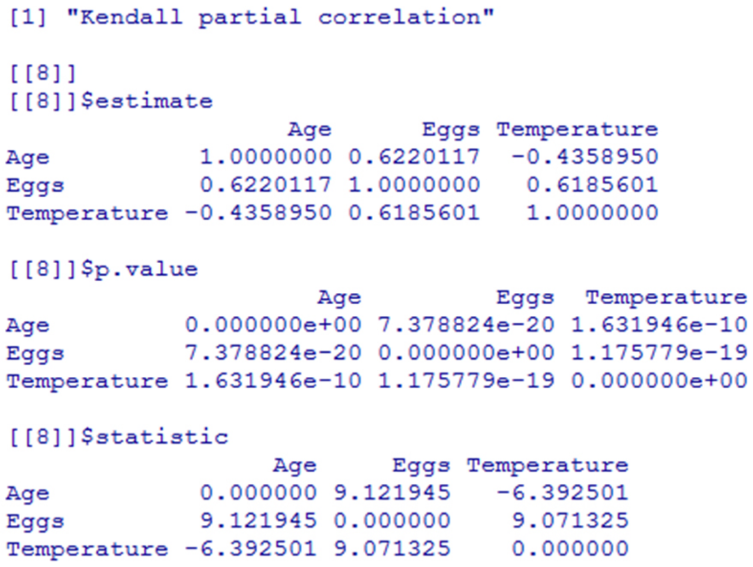


```
[1] "Pearson"

[[2]]

             Age    Eggs   Temperature
Age           *****  0.637 -0.107
Eggs         <0.001  *****  0.555
Temperature   0.290 <0.001  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values


[[3]]
[1] "Kendall"

[[4]]

             Age    Eggs   Temperature
Age           *****  0.498 -0.083
Eggs         <0.001  *****  0.493
Temperature   0.359 <0.001  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
```



```
[[5]]
[1] "Pearson partial correlation"

[[6]]
[[6]]$estimate
                   Age      Eggs Temperature
Age          1.0000000 0.8422472  -0.7185677
Eggs         0.8422472 1.0000000   0.8133817
Temperature -0.7185677 0.8133817   1.0000000

[[6]]$p.value
                     Age         Eggs  Temperature
Age         0.000000e+00 8.942973e-28 5.586822e-17
Eggs        8.942973e-28 0.000000e+00 1.489113e-24
Temperature 5.586822e-17 1.489113e-24 0.000000e+00

[[6]]$statistic
                  Age     Eggs Temperature
Age           0.00000 15.38732   -10.17615
Eggs         15.38732  0.00000    13.77078
Temperature -10.17615 13.77078     0.00000

[[6]]$n
[1] 100
```

Then the results including the variable temperature during the breeding season are shown. Again, only it should be noted in the ratio between the dependent variable and the independent, that for Pearson is 0.842 and 0.622 for Kendall. In both cases the probability, which is shown in the matrix *pvalue*, is significant for the two tests with p < 0.001. Therefore, as the two correlation coefficients increase when the control variable is not considered, temperature in this example, 0.637 to Pearson and 0.498 for Kendall, it was concluded that the relationship between the age of the female and the number of eggs per female varies depending on the temperature during the breeding season.



```
[1] "Kendall partial correlation"

[[8]]
[[8]]$estimate
                   Age      Eggs Temperature
Age          1.0000000 0.6220117  -0.4358950
Eggs         0.6220117 1.0000000   0.6185601
Temperature -0.4358950 0.6185601   1.0000000

[[8]]$p.value
                     Age         Eggs  Temperature
Age         0.000000e+00 7.378824e-20 1.631946e-10
Eggs        7.378824e-20 0.000000e+00 1.175779e-19
Temperature 1.631946e-10 1.175779e-19 0.000000e+00

[[8]]$statistic
                  Age     Eggs Temperature
Age          0.000000 9.121945   -6.392501
Eggs         9.121945 0.000000    9.071325
Temperature -6.392501 9.071325    0.000000
```

Value

A TXT file is obtained with the results of the correlations and can represent a graph relating the dependent variable with the independent, differentiating between the groups of the grouping variable.

References

Kim, S. (2013) Partial and Semi-partial (Part) correlation. R package version 1.0. Available at: http://CRAN.R-project.org/package=ppcor.

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, **17**, 1-25.

Rizopoulos, D. (2013) Latent Trait Models under IRT. R package version 1.0-0. Available at: http://CRAN.R-project.org/package=ltm.

Examples

```
## Not run:

data(ZX1)

X2(data=ZX1,varX="Age", varY="Eggs", varZ="Temperature")


## End(Not run)
```

X3 *MULTIPLE CORRELATION*

Description

A correlation analysis between all the variables selected by the user is applied and the probability values are also calculated. In addition, two graphics matrix of correlations are performed.

Usage

```
X3(data, variables, ResetPAR=TRUE, PAR=NULL, CORRPLOT=NULL, PAIRS=NULL,
colhist="#00FFFFFF", file="Output.txt")
```

Arguments

`data`	Data file.
`variables`	Variables for which the correlations are calculated.
`ResetPAR`	If FALSE, the default conditions of the function PAR are not placed and are those defined by the user in previous graphs.
`PAR`	It accesses the function PAR which allows to modify many different aspects of both charts.
`CORRPLOT`	It accesses the function corrplot which allows to specify the arguments of the first matrix chart.
`PAIRS`	It accesses the function pairs which allows to specify the arguments of the second matrix chart.
`colhist`	Colour of the bars of the histogram in the second graph matrix.
`file`	TXT FILE. Output file name with the results.

Details

X. CONTRASTS OF INDEPENDENCE AND ASSOCIATION IN QUANTITATIVE VARIABLES

X.3. MULTIPLE CORRELATION

X.3.1. Pearson's correlation coefficient

It is an extension of the method of simple Pearson correlation to cases where there is more than one independent variable. This coefficient indicates the degree of association between several independent variables x_i and a dependent variable *y*. This coefficient is often expressed as *R*, and its square R^2 is the "coefficient of determination", which is interpreted as the proportion of variance explained by the multiple linear regression:

$$R^2 = c' R_{xx}^{-1} c$$

where *c* it is a column vector or matrix with all possible correlations between the independent variables and the dependent variable. *c'* is the transposed in this array or row vector. R_{xx} is a matrix of correlations between all the independent variables, and the superscript -1 indicates that the inverse of this matrix is used in the formula. It can also be calculated as the correlation coefficient simple -bivariate- between the dependent variable and the estimate of *y* obtained using the multiple linear regression model, *y'*.

The significance can be obtained through the statistical *F*, calculated using the following formula:

$$F = \frac{\frac{R^2_{yx_{1...k}}}{k}}{\frac{(1-R^2_{yx_{1...k}})}{(n-k-1)}}$$

where the term $R^2_{yx_{1...k}}$ is the value of the coefficient of multiple correlation, *k* is the number of independent variables, and *n* the number of data. The degrees of freedom of the *F* are *k* and (*n* - *k* - 1).

X.3.2. Kendall's coefficient of concordance (W)

All the non-parametric coefficients to measure association between variables we have seen so far were applicable for bivariate data and are based on the comparison of two groups of ranges or ordinal variables. The Kendall's coefficient of concordance (*W*), however, will allow us to work with more than two groups of ranges or ordinal variables. It is therefore, to some extent, similar to the multiple correlation coefficient of Pearson, although in this case there is no a single dependent variable. Suppose that we have scores of *m* judges (for example tasters) on *n* objects (for example wines) and we want to measure the degree of agreement between judges. The formula to compute it is as follows

$$W = \frac{12S}{m^2(n^3 - n)}$$

where:

$$S = \sum_{i=1}^{n}(R_i - \bar{R})^2$$

R_i is the average range for the object *i* and $\bar{R}$ is the average of the ranges:

$$R_i = \sum_{j=1}^{m} r_{i,j} \quad \bar{R} = \tfrac{1}{2}m(n+1)$$

If "draws" between various data, each data is assigned to the mean of the ranks (if the third and fourth are equal, the range will be 3.5 for both).

To test the significance of this statistic, there would be two possible cases:

- If 3 $\leq N \leq$ 7, a distribution generated from all the possible combinations of ranges, which is tabulated for different number of variables or groups (3 $\leq k \leq$ 20), and large number of ranges (Siegel & Castellan, 1988) is used.

- If *N*>7, another statistical distribution that follows an approximately χ^2 with n - 1 degrees of freedom will be used:

$$\chi^2 = k(n-1)W$$

In any case, the H_0 that set, will be the value of *W* is equal to zero.

X.3.3. Squared coefficient of multiple linear correlation

Another method of calculation is the square of the coefficient of multiple linear correlation. That square is interpreted as the coefficient of determination or proportion of variance explained by the linear regression on the set of all of the explanatory variables.

The correlation matrix is calculated for the whole set, that array is reversed, and finally are subtracted from the unit the reciprocals of the elements of the diagonal, thus obtaining, simultaneously, all the coefficients of the squared multiple correlation. This is the method that will be used to estimate the correlation of one variable with multiple variables.

The square of the coefficient of multiple linear correlation is equal to the coefficient of determination r^2 which is obtained in the regressions.

FUNCTIONS

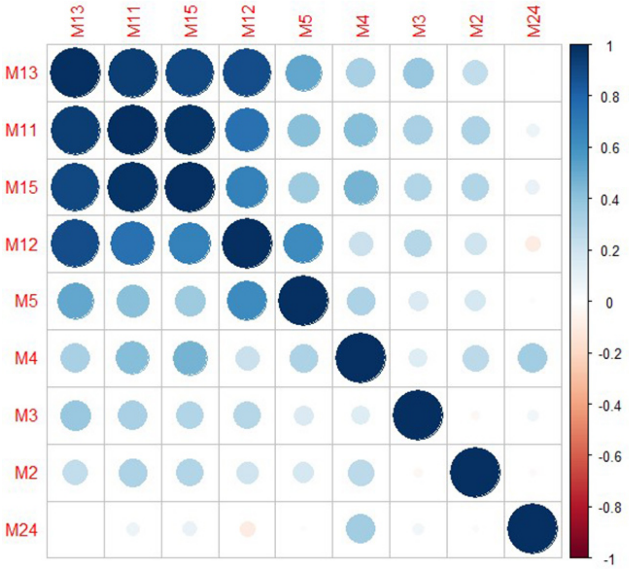
The function rcor.test of the package ltm (Rizopoulos 2006; 2013) for the calculation of the correlation and odds between the variables is used and the function smc of the package psych (Revelle, 2014) to calculate the multiple correlations. For the first graph the function corrplot of the package corrplot (Wei, 2013) is used. For the second graph functions panel.hist, panel.cor and panel.reg of SciViews package (Grosjean, 2014) and pairs of the base package graphics are used.

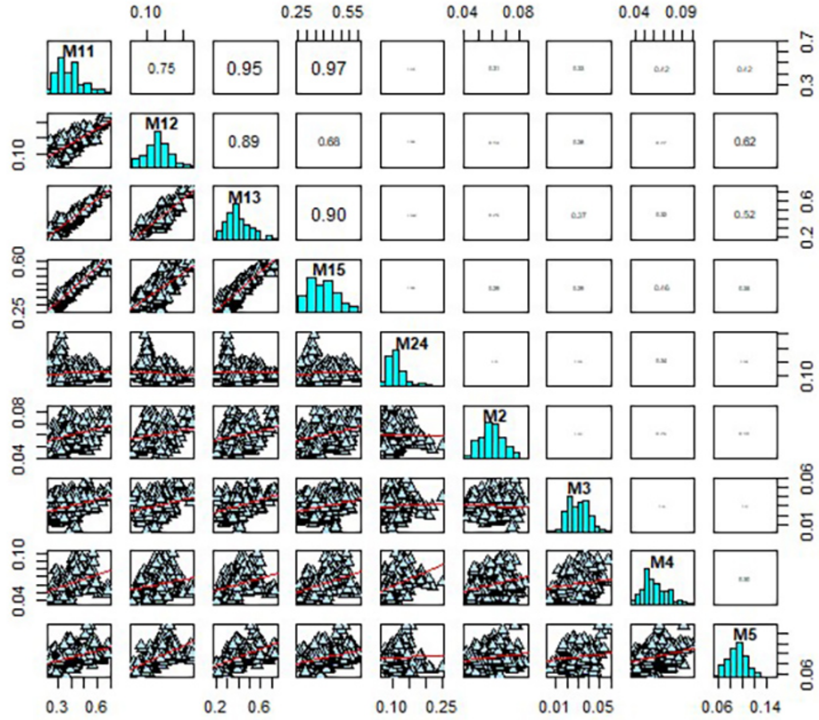
EXAMPLE

Morphometric data from various species of the order Characiformes, such as the length of the base of the dorsal fin (M12), height of the body (M11), and so on for more details see Guisande et al. (2010). The purpose is to estimate the correlation between several variables.

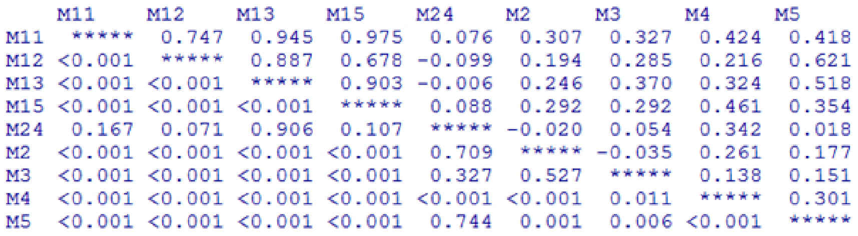
In Figures X.3 and X.4, the correlations between morphometric measurements of the fish in two different formats are represented. In the case of Figure X.3 variables are sorted by default using a Principal Components Analysis, although there are other methods available.

Figures X.3 y X.4 Correlation between morphometric measures of several fish species of the order Characiformes.



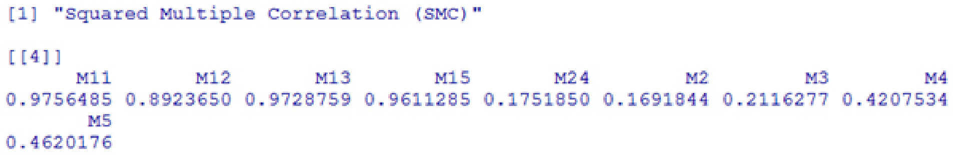


The results are shown in the following table, where the matrix above right are the correlations between variables, and the lower left the odds.



```
       M11     M12     M13     M15     M24     M2      M3      M4      M5
M11   *****   0.747   0.945   0.975   0.076   0.307   0.327   0.424   0.418
M12  <0.001   *****   0.887   0.678  -0.099   0.194   0.285   0.216   0.621
M13  <0.001  <0.001   *****   0.903  -0.006   0.246   0.370   0.324   0.518
M15  <0.001  <0.001  <0.001   *****   0.088   0.292   0.292   0.461   0.354
M24   0.167   0.071   0.906   0.107   *****  -0.020   0.054   0.342   0.018
M2   <0.001  <0.001  <0.001  <0.001   0.709   *****  -0.035   0.261   0.177
M3   <0.001  <0.001  <0.001  <0.001   0.327   0.527   *****   0.138   0.151
M4   <0.001  <0.001  <0.001  <0.001  <0.001  <0.001   0.011   *****   0.301
M5   <0.001  <0.001  <0.001  <0.001   0.744   0.001   0.006  <0.001   *****
```

Finally, in the following table, the multiple correlations of each variable with all the remaining ones are shown. For example, a value of 0.975 for the variable M11 means that the rest of variables explain 97.5% of the variability observed in M11.



```
[1] "Squared Multiple Correlation (SMC)"

[[4]]
      M11       M12       M13       M15       M24        M2        M3        M4 
0.9756485 0.8923650 0.9728759 0.9611285 0.1751850 0.1691844 0.2116277 0.4207534 
       M5 
0.4620176 
```

Value

A TXT file with the results of the correlations and probability values is obtained, in addition to performing two correlation matrix graphs.

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

Grosjean, P. (2014) SciViews GUI API - Main package. R package version 0.9-5. Available at: http://CRAN.R-project.org/package=SciViews.

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, **17(5)**, 1-25.

Rizopoulos, D. (2013) Latent Trait Models under IRT. R package version 1.0-0. Available at: http://CRAN.R-project.org/package=ltm.

Revelle, W. (2014) Procedures for Psychological, Psychometric, and Personality Research. R package version 1.4.5. Available at: http://CRAN.R-project.org/package=psych.

Siegel, S. & Castellan, N.J. Jr. (1988) *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Wei, T. (2013) Visualization of a correlation matrix). R package version 0.73. Available at: http://CRAN.R-project.org/package=corrplot.

Examples

```
## Not run:

data(ZX3)

#Correlation between morphometric variables of fish species #of order Characiformes

X3(data=ZX3, variables=c("M11","M12","M13","M15","M24","M2","M3","M4","M5"))


## End(Not run)
```

XI1 *SIMPLE REGRESSION*

Description

Different regression models are applied to describe the type of function that best fits the possible relationship between two variables.

Usage

```
XI1(data, varY, varX, model=c("Linear", "Log", "S-curve",
"Power", "Exp",
"Quadratic", "Cubic", "Inverse"), outliers=NULL, quant1=0.05, quant2 = 0.95,
ResetPAR=TRUE, mfrow=c(2,4), PAR=NULL, PLOT=NULL, PLOTR=NULL, YLAB=NULL,
XLAB=NULL, CEXPCH=1.3, COLABLINE="#48D1CCFF", COLOR="#C0FF3EFF",
PCH=15, resPlot=TRUE, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL,
file1="Output.txt", file2="Coefficients.csv", file3="Residuals.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data` Data file.

`varY` Dependent variable.

`varX` Independent variable.

`model` One, or multiple regression models can be selected: "Linear" , "Log", "S-curve", "Power", "Exp", "Quadratic", "Cubic", "Inverse". It is not considered the model in those cases in which there is the logarithm that apply to any of the variables, if any value of the variable, which applies the logarithm, is zero or negative. The inverse model is not calculated if any value of the independent variable is zero.

`outliers` If it is NULL, the outliers are not deleted. If any regression model, "Linear" , "Log", "S-curve", "Power", "Exp", "Quadratic", "Cubic" or "Inverse" is chosen, then the outliers are removed using the selected regression model.

`quant1` Quantile of the lower end to the elimination of outliers.

`quant2` Quantile of the upper end to the elimination of outliers.

`ResetPAR` If it is FALSE, the default condition of the function PAR is not placed and maintained those defined by the user in previous graphics.

`mfrow` If it is NULL and there are several regression models selected, these come in separate windows. If it is required that the graphs are displayed in panels, this argument is a vector with the format c(nr, nc) that indicates the number of figures per row (nr) and column (nc), by filling the first rows.

`PAR` It accesses the function PAR which allows to modify many different aspects of the graph.

`PLOT` It accesses the function plot.default which allows to modify the graph of the regression.

`PLOTR` It accesses the function plot.default which allows to modify the graph of the residuals.

`YLAB` Legend of the axis Y.

`XLAB` Legend of the axis X.

`CEXPCH` Size of the graphic symbols.

`COLABLINE` Color of the line of the regression model.

`COLOR` Colour of symbols.

`PCH` Type of symbol.

`resPlot` If TRUE, the graphs showing the relationship between the predicted value and the typified residual are shown.

`LEGEND` It allows to add a legend to the chart. It only makes sense if a single graph, since a single regression model is selected.

`AXIS` It allows to add axes to the graph. It only makes sense if a single graph, since a single regression model is selected.

`MTEXT` It allows to add text in the margins of the graph. It only makes sense if a single graph, since a single regression model is selected.

`TEXT` It allows to add text in any area of the inner part of the graph. It only makes sense if a single graph, since a single regression model is selected.

`file1` TXT FILE. Name of the output file with the results of the analysis.

`file2` CSV FILES. Filename with the coefficients of the regression models.

`file3` CSV FILES. Filename with residuals of the regression models.

`na` CSV FILES. Text that is used in the cells without data.

`dec` CSV FILES. It defines if the comma "," is used as decimal separator or the dot ".".

`row.names` CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

XI. REGRESSIONS

Regression is defined as the theory that seeks to express, through a mathematical function, the relationship between a dependent variable and a (simple regression) or multiple (multiple regression) independent variables. Obtaining this feature allows to predict what will be the value of the dependent variable depending on the value that the independent variable or variables take.

Regression differs from correlation in which the latter examines the degree of association between the variables, and determines whether the relationship is or is not significant, whereas the regression, as mentioned above, sought to define the role that best explains the relationship between the variables.

XI.1. REGRESSION MODELS FOR QUANTITATIVE DEPENDENT VARIABLES

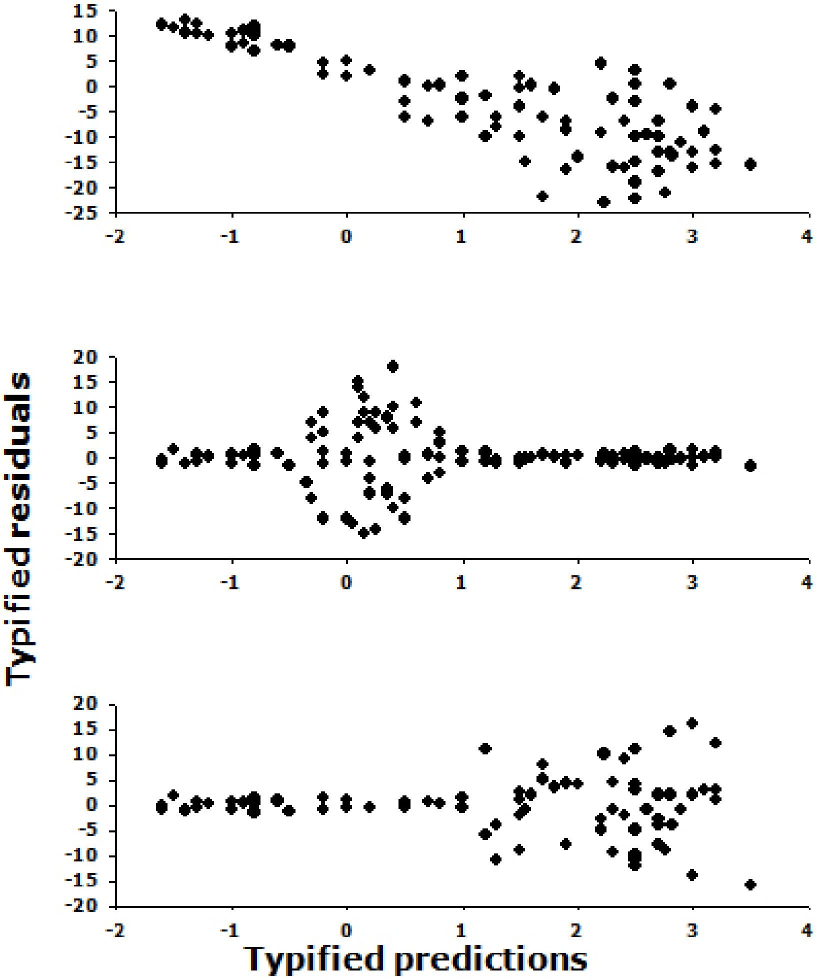
XI.1.1. Requirements

To implement a regression model between variables is not required that the data submitted a Normal distribution or that there is homogeneity of variances; however, in order to be able to determine if the function obtained with the regression model is significant it is necessary to apply contrasts, and this will have to fulfill the following requirements, which are largely related to the residuals (difference between the observed value of the dependent variable and the value set by the function):

1. The residuals obtained from the regression model must present a Normal distribution.

2. Homoscedasticity must exist in the residuals, that is to say, the variance of the residuals must be constant. The graphs shown below (Figure XI1.1) are examples where homoscedasticity of residuals is not satisfied, because its variability is not constant but changes depending on the predicted values. The first graph, obtained in the setting of a straight line, also shows that the relationship is not linear, since failure to maintain the horizontal trend.

Figure XI1.1. Examples without homoscedasticity.



3. There should be no autocorrelation in the series of residuals (must be independent). If this requirement is not met, it is not possible to know the exact degree of relationship between the independent variable and the dependent, because that part of the forecast is due to the own values of the dependent variable.

4. In the case of the multiple regression model, there must be no relationship between the independent variables, i.e. there should be no multicollinearity.

XI.1.2. Simple regression

The most common models, and that are calculated in this function, are the following, which are shown as they are expressed in the output file TXT file with the results of the models and the CSV where the coefficients of the models are displayed:

Model	**CSV**	**TXT**
Linear	$y = a + bx$	Same
Logarithm	$y = e^{a+\frac{b}{x}}$	Same

S-Curve	$y = e^{a+\frac{b}{x}}$	Same
Potential	$y = ax^b$	$y = e^{ln(a)+bln(x)}$
Exponential	$y = ae^{bx}$	$y = e^{ln(a)+bx}$
Quadratic	$y = a + bx + cx^2$	Same
Cubic	$y = a + bx + cx^2 + dx^3$	Same
Inverse	$y = a + \frac{b}{x}$	Same

FUNCTIONS

The function lillie.test of the package nortest (Gross, 2013) is used to perform the test of Normality Kolmogorov-Smirnov with Lilliefors'correction, the function dwtest of the package lmtest (Hothorn et al., 2013) to analyze the autocorrelation with the test and the Durbin-Watson statistic function bptest of the package lmtest (Hothorn et al., 2013) to perform the Breusch-Pagan test of homoscedasticity.

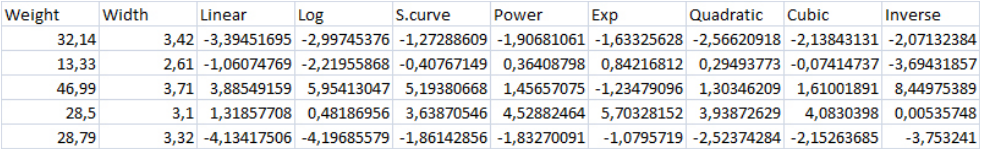
EXAMPLE

The data correspond to a study that aims to find the best regression model that exists between the weight and width of the shell for a set of three species of molluscs.

Step 1.

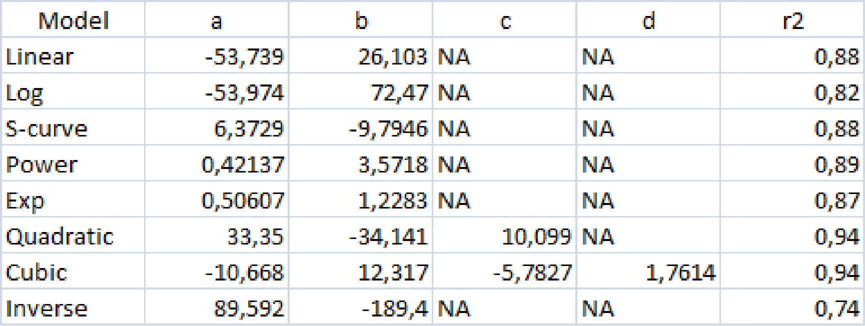
In the first place the relationship between the weight and width of the shell is analyzed considering the three species of molluscs and all of the regression models.

The residuals of the models, which are in one of the CSV files generated by this function are shown partially in the following table. These are useful to identify outliers, to see if it meets the homoscedasticity, as shown in the Figure XI1.1 and to analyze data for Normality.



Weight	Width	Linear	Log	S.curve	Power	Exp	Quadratic	Cubic	Inverse
32,14	3,42	-3,39451695	-2,99745376	-1,27288609	-1,90681061	-1,63325628	-2,56620918	-2,13843131	-2,07132384
13,33	2,61	-1,06074769	-2,21955868	-0,40767149	0,36408798	0,84216812	0,29493773	-0,07414737	-3,69431857
46,99	3,71	3,88549159	5,95413047	5,19380668	1,45657075	-1,23479096	1,30346209	1,61001891	8,44975389
28,5	3,1	1,31857708	0,48186956	3,63870546	4,52882464	5,70328152	3,93872629	4,0830398	0,00535748
28,79	3,32	-4,13417506	-4,19685579	-1,86142856	-1,83270091	-1,0795719	-2,52374284	-2,15263685	-3,753241

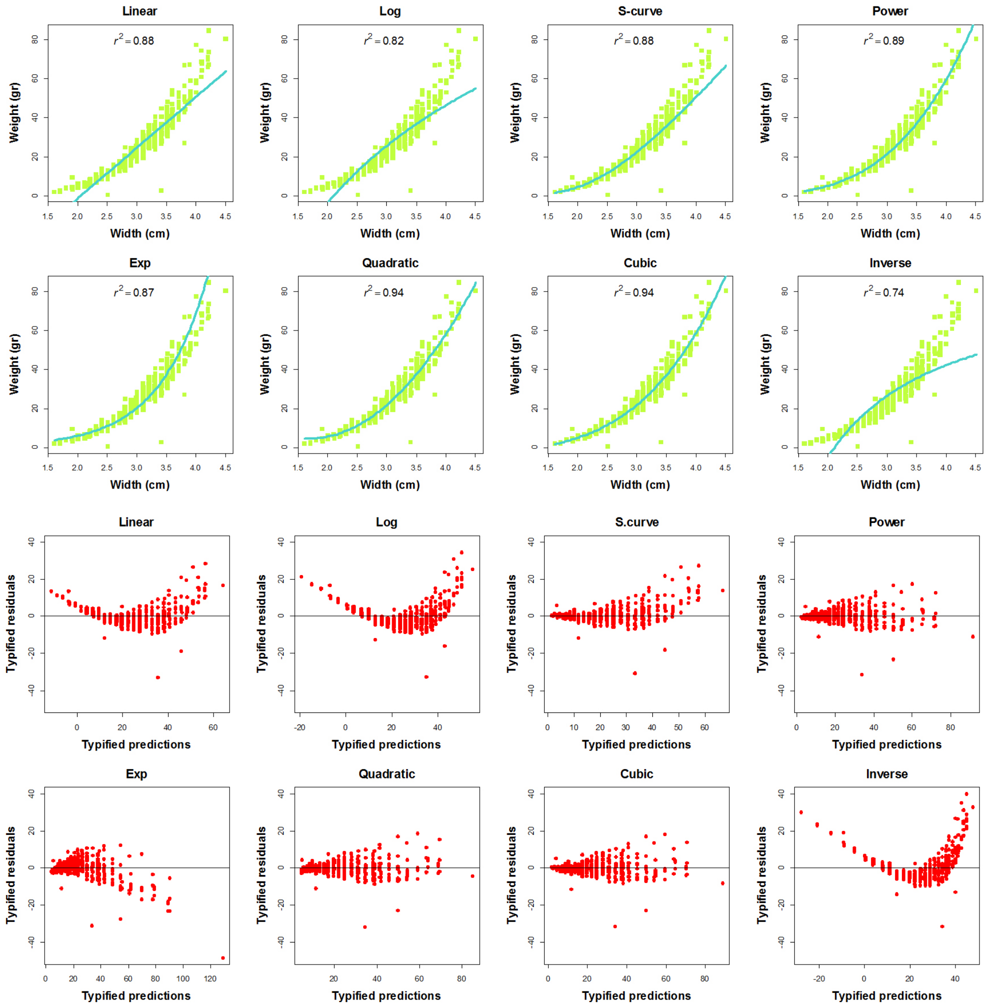
The following table shows the coefficients of the equations of regression models and r^2.



Model	a	b	c	d	r2
Linear	-53,739	26,103	NA	NA	0,88
Log	-53,974	72,47	NA	NA	0,82
S-curve	6,3729	-9,7946	NA	NA	0,88
Power	0,42137	3,5718	NA	NA	0,89
Exp	0,50607	1,2283	NA	NA	0,87
Quadratic	33,35	-34,141	10,099	NA	0,94
Cubic	-10,668	12,317	-5,7827	1,7614	0,94
Inverse	89,592	-189,4	NA	NA	0,74

Figure XI1.2, which generates the function, shows the eight regression models in a single panel and in Figure XI1.3 the residuals (the black line indicates a residual of zero).

Figures XI1.2 and XI1.3. Relationship between shell width and weight in three species of molluscs and residuals of the regressions.



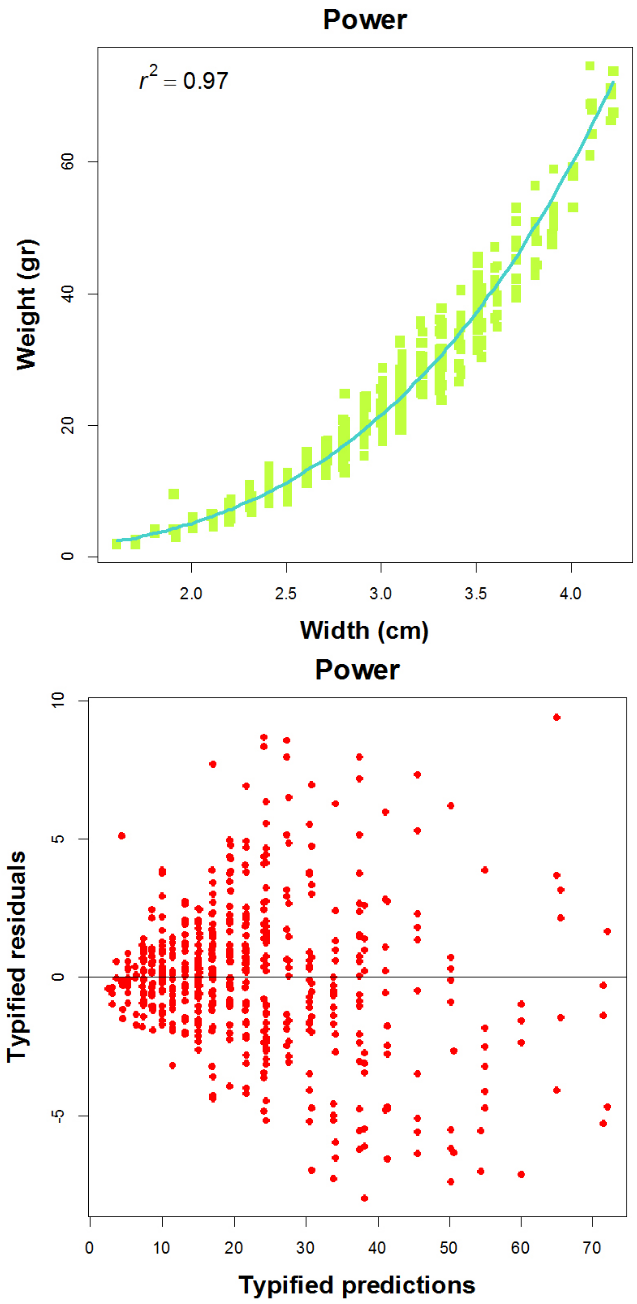
Although the values of r^2 are higher in the quadratic and cubic models, it is observed that the potential model residuals are equally small, have a linear behavior and are very similar throughout the entire range of predicted values of the dependent variable, to those obtained with quadratic and cubic models (Figure XI1.3). Therefore, the potential model was chosen because it is preferable to choose a simpler model, provided that the residuals are similar to those obtained in the quadratic and cubic models.

Step 2.

In the figure XI1.2 it was noted that there were some outliers, which can be removed easily using the argument *outliers*. As mentioned before that the best model was the potential, it is now defined with the argument *model= "Power"*. In addition, it specifies that data is removed using the same model *outliers= "Power"* and as only a graph is removed, the panel is canceled with *mfrow=NULL*.

The new model that is obtained (figure XI1.4) shows that the outliers have disappeared and due to this the $r^2 = 0.97$, is greater than in Figure XI 1.2 (r^2 = 0.89) where the atypical data had not been deleted. Figure XI1.5 shows the residuals, which then will be whether they are homogeneous throughout the range of predicted values of the dependent variable.

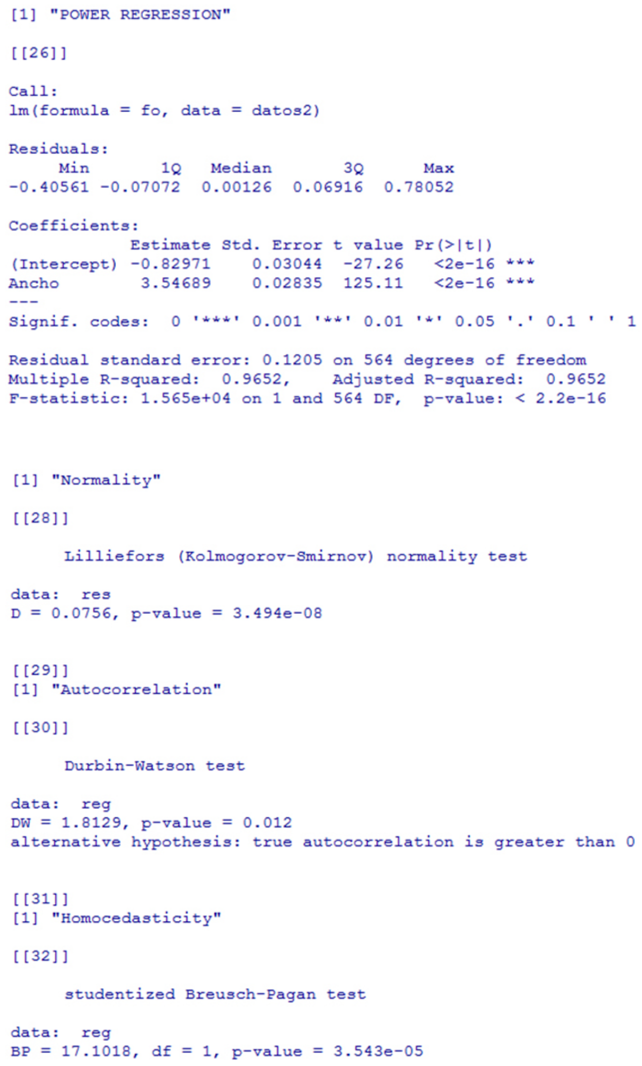
Figures XI1.4 and XI1.5 Potential relationship between shell width and weight and between the predicted values and residuals.



In the TXT file that generates the function, the regression model power is shown, where the variable shell width is significant (p < 0.001, see *Pr(>|t|)*) and, therefore, the model as a whole was also significant (p < 0.001, see *p-value* at the end of the results).

The r^2 (see *Multiple R-squared*) shows that the width of the shell explains a 97% of the observed variance in the weight of the individuals. The r^2 adjusted (see *Ajusted R-squared*) takes into account the size of the sample to determine the proportion above and, in this case, it shows the same value. The r^2 adjusted should be used to compare models with different numbers of observations or independent variables. The equation of the potential regression model could be expressed in either of these two ways:

$$Weight = 0.436 * Width^{3.547} \quad \text{or} \quad Weight = e^{ln(0.436)+3.547*ln(Width)}$$



```
[1] "POWER REGRESSION"

[[26]]

Call:
lm(formula = fo, data = datos2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.40561 -0.07072  0.00126  0.06916  0.78052 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.82971    0.03044  -27.26   <2e-16 ***
Ancho        3.54689    0.02835  125.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1205 on 564 degrees of freedom
Multiple R-squared:  0.9652,    Adjusted R-squared:  0.9652 
F-statistic: 1.565e+04 on 1 and 564 DF,  p-value: < 2.2e-16



[1] "Normality"

[[28]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.0756, p-value = 3.494e-08


[[29]]
[1] "Autocorrelation"

[[30]]

     Durbin-Watson test

data:  reg
DW = 1.8129, p-value = 0.012
alternative hypothesis: true autocorrelation is greater than 0


[[31]]
[1] "Homocedasticity"

[[32]]

     studentized Breusch-Pagan test

data:  reg
BP = 17.1018, df = 1, p-value = 3.543e-05
```

In the above table, the results of the test of Kolmogorov-Smirnov normality with Lilliefors' correction are shown, the test for autocorrelation of Durbin-Watson statistic and the Breusch-Pagan test of homoscedasticity.

Normality The residuals do not have a Normal distribution with p < 0.001. Although is not complied with the assumption of normality, this does not invalidate the model as it is very predictive with a r^2 very high. The problem resulting from these residuals are not Normal is that there can be no assurance that the degree of significance, probability value that shows the model, is the correct one.

Autocorrelation The requirement that there should be no autocorrelation is no longer met the test of Durbin-Watson statistic p = 0.012. This means that the value of r^2 of the 97% is not all due to the independent variable, the width of the shell, but it is also in part due to the own dependent variable that is auto explained and, therefore, it is not possible to know exactly how much is the variance explained by the independent variable. Anyway it is necessary to mention that the probability value of the test of Durbin-Watson statistic can be less than 0.05 easily when there are many data, as in this example. The statistical DW, whose value is 1.81 in this example, is a better indicator of the autocorrelation when the number of data is very large. According to Durbin & Watson (1951), a DW less than 1 means a strong positive autocorrelation, a DW greater than 4 a strong negative autocorrelation, values between 1 and 3 a moderate autocorrelation, and a value close to 2 means that there is no autocorrelation. Therefore, it can be concluded that the autocorrelation in this example is very small and the variance explained by the independent variable may be close to 97%.

Homoscedasticity Finally, the requirement of homoscedasticity of the residuals is not satisfied, because the likelihood of the Breusch-Pagan test is p < 0.001. The fact of not fulfilled this requirement means that the model is not as predictive for the entire range of values of the dependent variable. In the residuals of the model (Figure XI1.5) it is noted that as the width is greater, the residuals are larger. Therefore, the model predicts well the weight in smaller individuals, and a little worse in the mollusks of greater size.

Value

A TXT file is obtained with the results of the regressions and the normality test, homogeneity of variances and homoscedasticity for each of the regression models, as well as graphics with the relationship between the dependent and independent variable for each of the regression models, in addition to graphics with the relationship between the predicted value and the typified residual.

References

Durbin, J. & Watson G.S. (1951) Testing for serial correlation in least squares regression. *Biometrika*, **38**, 159-171.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Hothorn, T. et al., (2013) Testing Linear Regression Models R package version 0.9-33. Available at: http://CRAN.R-project.org/package=lmtest.

Examples

```
## Not run:

data(ZXI1)

#Step 1. Regression between shell width (independent variable) and
#weight (dependent variable) considering several species of molluscs
```

```
XI1(data=ZXI1, varX="Width",varY="Weight", XLAB= "Width (cm)", YLAB="Weight (gr)")

#Step 2. Same regression but removing outliers and performing
#only potential model

XI1(data=ZXI1, varX="Width",varY="Weight", model="Power", outliers="Power",
mfrow=NULL, XLAB= "Width (cm)", YLAB="Weight (gr)")


## End(Not run)
```

XI10 *POISSON REGRESSION-ESTIMATION*

Description

A Poisson regression is applied which allows to find the model that best fits the possible relationship between the number of times it occurs when regarding some random phenomenon (dependent variable) and several independent variables.

Usage

```
XI10(data, varY, varX, varplot=NULL, offset=NULL, offsetlog=TRUE, stepwise=TRUE,
ResetPAR=TRUE, PAR=NULL,  YLAB=NULL, XLAB=NULL, PLOT.TS=NULL, OrderCat=NULL,
LabelCat=NULL, COLOR=NULL, PCH=NULL, LEGEND=NULL, AXIS=NULL, MTEXT=NULL,
TEXT=NULL, file1="Coefficients.csv", file2="Predictions.csv", na="NA", dec=",",
row.names=FALSE, file3="Output.txt")
```

Arguments

`data` Data file.

`varY` Dependent variable.

`varX` Independent variables.

`varplot` Vector with three variables. The first variable is the Y axis, the second the X axis and the third is like a factor, in such a way that the relationship between the first and the second variable is displayed on each of the different categories coming from the third variable of this vector *varplot*.

`offset` A weight variable can be specified.

`offsetlog` If it is TRUE, the variable *offset* is logarithmically transformed.

`stepwise` If it is TRUE, regression is applied according to steps in order to eliminate those variables that are not significant. It uses the Akaike Information Criterion (*AIC*)

.

`ResetPAR` If FALSE, the conditions are not placed by default in the function PAR and those defined by the user on previous graphics are maintained.

`PAR`	It accesses the function PAR that allows to modify many different aspects of the graph.
`YLAB`	Legend of the Y axis.
`XLAB`	Legend of the X axis.
`PLOT.TS`	It accesses the function plot.ts from the stats package.
`OrderCat`	It allows to specify a vector with the order in which the categories from the X axis are shown.
`LabelCat`	It allows to specify a vector with the names of the categories from the X axis.
`COLOR`	Colors that are used in the graph time series, to identify the categories of the third variable which is specified in the argument of the vector *varplot*.
`PCH`	Symbols that are used in the time series graph, to identify the categories of the third variable which is specified in the argument vector *varplot*.
`LEGEND`	It allows to include a legend.
`AXIS`	It allows to add axes.
`MTEXT`	It allows to add text in the margins of the graph.
`TEXT`	It allows to add text in any area of the inside part of the graph.
`file1`	CSV FILES. Filename with the coefficients of the Poisson regression.
`file2`	CSV FILES. Filename with the predictions of the Poisson regression.
`na`	CSV FILES. Text used in the cells without data.
`dec`	CSV FILES. It defines if the comma "," is used as decimal separator or the dot ".".
`row.names`	CSV FILES. Logical value that specifies whether identifiers are being placed in rows or either a vector with a text for each one of the rows.
`file3`	TXT FILE. Output file name with the results of the analysis.

Details

XI. REGRESSIONS

XI.1. REGRESSION MODELS FOR QUANTITATIVE DEPENDENT VARIABLES

XI.1.6. Poisson regression

XI.1.6.1. Model estimation

It is used to model the number of times that a true random phenomenon occurs. For example, the number of new cases of a particular disease during a given time period in fairly general terms it can be adequately modeled by a Poisson distribution. It suffices to do with the assumption that cases occur randomly over time and independently. This is true for many other series of events, like the calls to a telephone switchboard machine breakdowns, or traffic accidents, but in the epidemiological field the number of new cases is used to measure the incidence rate, so the Poisson distribution turns out to be important in the study of the incidence and its causes.

When using one or more explanatory factors, that divide the population into strata (smokers and nonsmokers, groups receiving different medication, age groups, etc), a model is usually built to try to explain the differences between groups, admitting that within each group variability occurs randomly.

The Poisson distribution is simple: it has only one parameter, generally represented by λ. This value is both the mean and variance of the variable at the same time, so their estimate is very simple: it is enough to just take a representative sample and calculate the average. If in a particular area 7.2 of the cases annually appear as an average disease (average value, for example, in the last 10 years), that is the estimated value of the lambda parameter.

If in the sample we observe the variance being very different from the average, that will indicate us that the Poisson distribution is not adequate to model that situation, but overall the Poisson regression is appropriate for variables that measure recounts, in which case the linear regression is not appropriate, since the estimate could produce negative recounts, residuals does not have a normal distribution, and the variance of the response tends to increase with the average.

In the Poisson regression model it is assumed that the dependent variable, given the values of the explanatory variables, has a Poisson distribution (conditional distribution by those values) whose parameter λ has the value:

$$\lambda = e^{P(X_i)}$$

being $P(X_i)$ of a linear function of the explanatory variables.

FUNCTIONS

To estimate the Poisson regression, the function used is the one glm from the stats package.

To estimate the model by steps the function used stepAIC was from the MASS package (Venables & Ripley, 2002; Ripley et al., 2014).

The time series graph is done with the function plot.ts from the stats package.

The contrast to determine if it is a Poisson distribution is being performed with the old function *epicalc*.

EXAMPLE

Data refer to a study of cohorts, about the exposure of arsenic in the industry and the deaths from respiratory diseases, available in the database Montana from the old emphepicalc package (Chongsuvivatwong, 2012).

The data is the number of deaths in each group (*respdeath*), number of persons by years of exposure (*personyrs*), dividing the number of deaths and the number of persons per years of exposure * 10.000 (*respdeath.personyrs*), age groups (*agegr*), period of employment (*period*), start of employment (*start*) and time of exposure to arsenic throughout the years (*arsenic*).

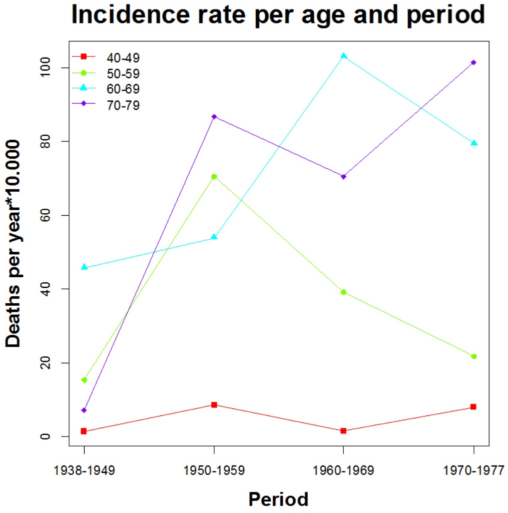
The primary endpoint of this study is (*respdeath*), number of deaths from respiratory causes, and we try to find the relationship that may exist between this variable and the others remaining. Since the target variable is a recount, the Poisson model may be suitable. For the regression model the dependent variable is the number of deaths in each group (*varY=respdeath*).

The independents are age group (*agegr*), period of employment (*period*), start of the employment (*start*) and time of exposure to the arsenic throughout the years (*arsenic*), which are defined on the argument *varX=c("agegr", "period", "start", "arsenic")*.

The argument *offset* allows to include a constant or a variable in the function. As we know a priori that the number of deaths should be proportional in each group to the number of people exposed, in the example the variable *personyrs*, we include this variable in the model, being necessary to transform the variable logarithmically so it can correspond with the overall transformation performed with the remaining variables, for which we leave the default argument *offsetlog=TRUE*.

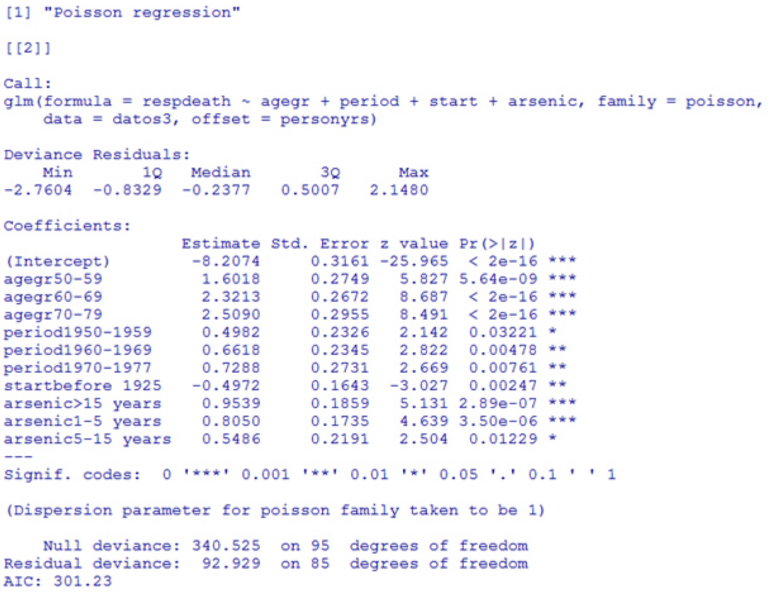
With the argument *varplot=c("respdeath.personyrs","period","agegr")* variables are defined to represent the time series graph.

Figure XI11.1. Evolution of the incidence rate (deaths divided by number of exposed people) for different periods of work and different age groups.



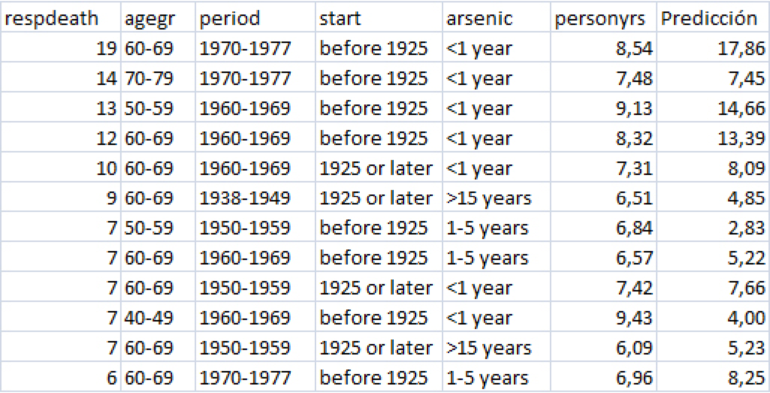
In the graph (Figure XI11.1) it is observed that in the incidence, number of deaths divided by the number of people who were exposed, is higher in the two older age groups.

The following results show the model obtained, whose $r^2 = 0.8$ and the contrast, whose null hypothesis is that the variable follows a Poisson distribution, has a p = 0.26. Therefore, we conclude that the distribution is of Poisson.



```
[1] "Poisson regression"

[[2]]

Call:
glm(formula = respdeath ~ agegr + period + start + arsenic, family = poisson,
    data = datos3, offset = personyrs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7604  -0.8329  -0.2377   0.5007   2.1480

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -8.2074     0.3161 -25.965  < 2e-16 ***
agegr50-59          1.6018     0.2749   5.827 5.64e-09 ***
agegr60-69          2.3213     0.2672   8.687  < 2e-16 ***
agegr70-79          2.5090     0.2955   8.491  < 2e-16 ***
period1950-1959     0.4982     0.2326   2.142  0.03221 *
period1960-1969     0.6618     0.2345   2.822  0.00478 **
period1970-1977     0.7288     0.2731   2.669  0.00761 **
startbefore 1925   -0.4972     0.1643  -3.027  0.00247 **
arsenic>15 years    0.9539     0.1859   5.131 2.89e-07 ***
arsenic1-5 years    0.8050     0.1735   4.639 3.50e-06 ***
arsenic5-15 years   0.5486     0.2191   2.504  0.01229 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 340.525  on 95  degrees of freedom
Residual deviance:  92.929  on 85  degrees of freedom
AIC: 301.23
```

The following table shows the predictions for each case that allow to analyze the relative risk.



respdeath	agegr	period	start	arsenic	personyrs	Predicción
19	60-69	1970-1977	before 1925	<1 year	8,54	17,86
14	70-79	1970-1977	before 1925	<1 year	7,48	7,45
13	50-59	1960-1969	before 1925	<1 year	9,13	14,66
12	60-69	1960-1969	before 1925	<1 year	8,32	13,39
10	60-69	1960-1969	1925 or later	<1 year	7,31	8,09
9	60-69	1938-1949	1925 or later	>15 years	6,51	4,85
7	50-59	1950-1959	before 1925	1-5 years	6,84	2,83
7	60-69	1960-1969	before 1925	1-5 years	6,57	5,22
7	60-69	1950-1959	1925 or later	<1 year	7,42	7,66
7	40-49	1960-1969	before 1925	<1 year	9,43	4,00
7	60-69	1950-1959	1925 or later	>15 years	6,09	5,23
6	60-69	1970-1977	before 1925	1-5 years	6,96	8,25

Value

A TXT file is obtained with the regression results, two CSV files with the coefficients and regression predictions, and a time series chart if selected.

References

Chongsuvivatwong, V. (2012) Epidemiological calculator. R package version 2.15.1.0.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: http://CRAN.R-project.org/package=MASS.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXI11)

XI10(data = ZXI11, varY = "respdeath", varX = c("agegr", "period",
"start", "arsenic"), offset = "personyrs", varplot = c("respdeath.personyrs",
"period", "agegr"), YLAB = "Deaths per year*10.000", XLAB = "Period",
MTEXT = c("text = 'Incidence rate per age and period'", "line = 1", "cex = 2",
"font = 2"))


## End(Not run)
```

XI11 *POISSON REGRESSION-PREDICTION*

Description

A Poisson regression model previously calculated with XI10 function is applied.

Usage

```
XI11(data, varY,  varX, offset=NULL,  offsetlog=TRUE, file="Model predictions.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Data file.
`varY`	Dependent variable.
`varX`	Independent variables.
`offset`	A weight variable may be specified.
`offsetlog`	If it is TRUE the variable *offset* is then logarithmically transformed.
`file`	CSV FILES. Filename with the predictions of the Poisson regression.
`na`	CSV FILES. Text used in the cells without data.
`dec`	CSV FILES. Defines if a decimal separator like a comma "," or period "." is being used.
`row.names`	CSV FILES. Logical value that specifies whether identifiers are put in rows or in a vector with a text for each of the rows set.

Details

XI. REGRESSIONS

XI.1. REGRESSION MODELS FOR QUANTITATIVE DEPENDENT VARIABLES

XI.1.6. Poisson Regression

XI.1.6.2. Model prediction

The best method to assess the quality of the regression model is to check the degree of accuracy using data that was not included in the original matrix and which was used to estimate the model. This function allows to easily apply a Poisson regression model already estimated which, besides serving to evaluate the quality of the model, it also serves to be able to use the model, i.e., predict values predicted for a given case.

To use this feature, it will be necessary to have the Poisson regression model calculated with the function XI10.

EXAMPLE

The data must have the same format, i.e., the same variables and the same position, that the data matrix which was used when the model was estimated with the function XI10. In addition, the working directory must be the same that was used just like when the model was estimated using the function XI10.

Data refer to a study of cohorts, about the exposure of arsenic in the industry and the deaths from respiratory diseases, available in the database of the old package *epicalc* from the epicalc package (Chongsuvivatwong, 2012). The data is the number of deaths in each group (*respdeath*), number of persons by years of exposure (*personyrs*), dividing the number of deaths and the number of persons per years of exposure * 10.000 (*respdeath.personyrs*), age groups (*agegr*), period of employment (*period*), start of employment (*start*) and time of exposure to arsenic throughout the years (*arsenic*).

However, these data were not included when the original model was estimated with the XI10 function. The primary endpoint of this study is (*respdeath*), number of deaths from respiratory causes, we try to find the relationship that may exist between this variable and the others remaining. Since the target variable is a recount, the Poisson model may be suitable.

For the regression model the dependent variable is the number of deaths in each group (*varInteresY=respdeath*). The independents are age groups (*agegr*), period of employment (*period*), start of employment (*start*) and time of exposure to arsenic in years (*arsenic*), which are defined on the argument *varInteresX=c("agegr", "period", "start", "arsenic")*.

The argument *offset* allows to include a constant or a variable in the function. As we know a priori, the number of deaths should be proportional in each group as the number of people exposed, in the example, the variable *personyrs*, we include this variable in the model, being necessary for the variable logarithmically transformed to correspond with the overall transformation performed with the remaining variables, for which leave the default argument intact *offsetlog=TRUE*.

The results show that $r^2 = 0.84$. Therefore, it seems to be that the model is very trustworthy and reliable.

Value

A CSV file is obtained with the model predictions.

Examples

```
## Not run:

data(ZXI12)

XI11(data=ZXI12, varY="respdeath", varX=c("agegr", "period", "start",
"arsenic"), offset="personyrs")


## End(Not run)
```

XI2 *NON-LINEAR FUNCTIONS*

Description

Different regression models are applied to describe the type of non-linear function that best fits the possible relationship between two variables.

Usage

```
XI2(data, varY, varX, model=c("Clench", "Nexponential", "Saturation",
"Rational", "Bertalanffy", "Gompertz", "Logistic"), outliers=NULL,
quant1=0.05, quant2 = 0.95, ResetPAR=TRUE, mfrow=c(2,4), PAR=NULL,
PLOT=NULL, YLAB=NULL, XLAB=NULL, CEXPCH=1.3, COLABLINE="#48D1CCFF",
COLOR="#C0FF3EFF", PCH=15, resPlot=TRUE, LEGEND=NULL, AXIS=NULL, MTEXT= NULL,
TEXT=NULL, file1="Output.txt", file2="Coefficients.csv", file3="Residuals.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data` Data file.

`varY` Dependent variable.

`varX` Independent variable.

`model` One or multiple regression models can be selected: "Clench", "Nexponential", "Saturation", "Rational", "Bertalanffy", "Gompertz" and/or "Logistic".

`outliers` If NULL, outliers are not removed. When choosing a regression model, "Clench", "Nexponential", "Saturation", "Rational", "Bertalanffy", "Gompertz" or "Logistic", then the outliers are removed using the selected regression model (for further information see section *details*).

`quant1` Quantile of the lower end to the elimination of outliers.

`quant2` Quantile of the upper end to the elimination of outliers.

`ResetPAR` If it is FALSE, the default condition of the function PAR is not placed and maintained those defined by the user in previous graphics.

`mfrow` If it is NULL and there are several regression models selected, these come in separate windows. If it is required that the graphs are displayed in panels, this argument is a vector with the format c(nr, nc) that indicates the number of figures per row (nr) and column (nc), by filling the first rows.

`PAR` It accesses the function PAR which allows to modify many different aspects of the graph.

`PLOT` It accesses the function plot.default which allows to modify the graph of the regression.

`YLAB` Legend of the axis Y.

`XLAB` Legend of the axis X.

`CEXPCH` Size of the graphic symbols.

`COLABLINE` Color of the line of the regression model.

`COLOR` Colour symbols.

`PCH` Type of symbol.

`resPlot` If TRUE, the graphs showing the relationship between the predicted value and the typified residual are shown.

`LEGEND` It allows to add a legend to the chart. It only makes sense if a single graph, since a single regression model is selected.

`AXIS` It allows to add axes to the graph. It only makes sense if a single graph, since a single regression model is selected.

`MTEXT` It allows to add text in the margins of the graph. It only makes sense if a single graph, since a single regression model is selected.

`TEXT` It allows to add text in any area of the inner part of the graph. It only makes sense if a single graph, since a single regression model is selected.

`file1` TXT FILE. Name of the output file with the results of the analysis.

`file2` CSV FILES. Filename with the coefficients of the regression models.

`file3` CSV FILES. Filename with residual of the regression models.

`na` CSV FILES. Text that is used in the cells without data.

`dec` CSV FILES. It defines if the comma "," is used as decimal separator or the dot ".".

`row.names` CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

XI. REGRESSIONS

XI.1. REGRESSION MODELS FOR QUANTITATIVE DEPENDENT VARIABLES

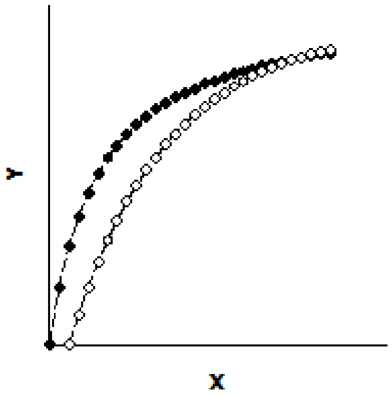
XI.1.3. Non-linear functions

Many times variables are adjusted to a type of non-linear function that is none of the common which are described in most statistical programs as were observed in the function XI1. Below are some of these non-linear functions and how to calculate them.

XI.1.3.1. Accumulation functions

The growth rate of a bacterial population, of yeast or phytoplankton at different concentrations of nutrients or temperature, the photosynthetic rate of plants at different light intensities, the accumulated richness in function of the number of samplings, among many other examples, are adjusted to the so-called functions of accumulation (Figure XI2.1). There are different types but they can be divided between those in which the zero of Y and the zero of X always coincide and those that allow a non-zero value for Y when X is zero (Figure XI2.1).

Figure XI2.1. Accumulation curves in which zero of Y and zero of X always coincide (black circles) and curves that allow a different value of zero for X when Y is zero (white circles).



In this *script* is calculated: 1) the curve of Clench (Clench, 1979), which is a modification of the function of Monod (Monod, 1950), and was proposed to butterflies; 2) the negative exponential (Miller & Wiegert, 1989) that was proposed for rare plant species; 3) the saturation curve that was used to show the relationship between growth of phytoplankton, a toxic algae of the genus *Alexandrium* and the concentration of phosphate (Frangópulos et al., 2004), which is similar to von Bertalanffy growth curve that we will see below but adapting the coefficients to better explain the pattern of accumulation function and, finally; 4) the rational function (Ratkowski, 1990) that can be used when there is no clear criterion which model to use (Falther, 1996).

Name	**Function**	**Reference**
Clench	$y = \frac{ax}{1+bx}$	(Clench, 1979)
Negative exponential	$y = a\left(1 - e^{-bx}\right)$	(Miller & Wiegert, 1989)
Saturation	$y = a\left(1 - e^{-b(x-c)}\right)$	(Frangópulos et al., 2004)
Rational	$y = \frac{(a+bx)}{(1+cx)}$	(Ratkowski, 1990)

XI.1.3.2. Growth functions

The growth curves are generally applied to the growth of individuals while the von Bertalanffy curve can also be set to the growth of populations. That is why previously showed the saturation curve, whose formula is identical to the von Bertalanffy, but with their coefficients modified.

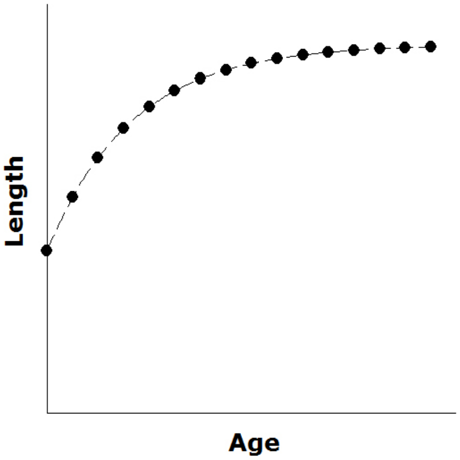
XI.1.3.2.1. Growth curve of von Bertalanffy

The growth curve of many organisms fits well with the von Bertalanffy equation (Figure XI2.2):

$$L_t = L_\infty \left(1 - e^{-k(t-t_0)}\right)$$

where L_t is the length of the individual to the age *t*, L_∞ is the maximum length that reaches the individual when the growth ceases, *k* is the constant growth expressed in $time^{-1}$ and t_0 is the hypothetical age that would have an individual who had a zero size. Therefore, the dependent variable is L_t, the independent variable is the age of the individual (*t*) and the constants in this equation are *k*, L_∞ and t_0.

Figure XI2.2. Growth curve of von Bertalanffy



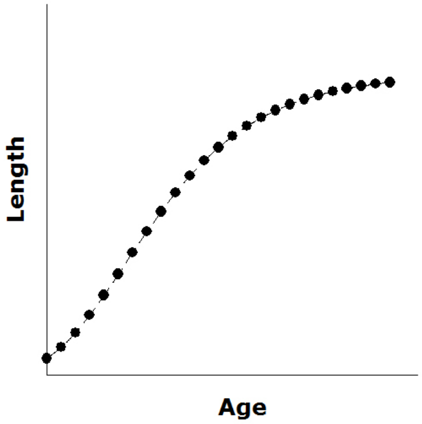
XI.1.3.2.2. Growth curve of Gompertz

In some species, such as fish, crustaceans and molluscs, in the early stages of the life cycle (larval stage) the growth can be better adjusted to the Gompertz equation (Hernandez-Llamas & Ratkowsky, 2004) (Figure XI2.3):

$$L_t = ae^{-be^{-ct}}$$

where L_t is the length of the individual to the age *t*, *a* is a constant that represents the maximum value of the asymptotic curve, *t* is the age of the individual and, *b* and *c* are the other two constants in the equation.

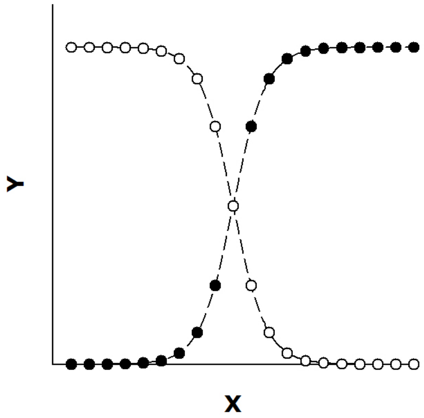
Figure XI2.3. Growth curve of Gompertz



XI.1.3.2. Logistics function

Many phenomena in nature are adjusted to a logistic curve (Figure XI2.4), as for example the changes in the abundance of a population over time (Smith & Smith, 2000), the scientific progress over time within a particular line of research (Solla Price, Little Science, Big Science, 1963 -cited in Callon et al., 1995), etc.

Figure XI2.4. Logistics function with a negative slope (white circles) and positive (black circles).



The logistic curve is defined by the following equation:

$$Y = \frac{a}{1 + e^{(b-cX)}}$$

where the constant *a* sets the upper limit of the curve, and is equal to *K* or load capacity of the population (maximum number of individuals of that population), in the case of that model a population.

The constant c determines the slope, and in the case of populations is equal to r or rate of population growth. The constant b defines the size of the so-called phase of latency in the case of populations, the phase with smaller values before starting the slope to reach the maximum value.

FUNCTIONS

It uses the function lillie.test of the package (Gross, 2013) to perform the test Kolmogorov-Smirnov normality with Lilliefors' correction and the function nls of the base package stats for the estimation of the models.

EXAMPLE 1

Samplings were conducted at several sites and the number of species present was recorded. It was a relationship between the number of sites sampled and accumulated richness considering all sites surveyed above. The goal is to determine the theoretical total number of species that exists in the area.

Figure XI2.5 shows the functions obtained, where it is observed that the functions of Clench and Rational are those that have the higher r^2.

Figure XI2.5. Relationship between the number of sites sampled and richness of accumulated species considering all sites surveyed so far.

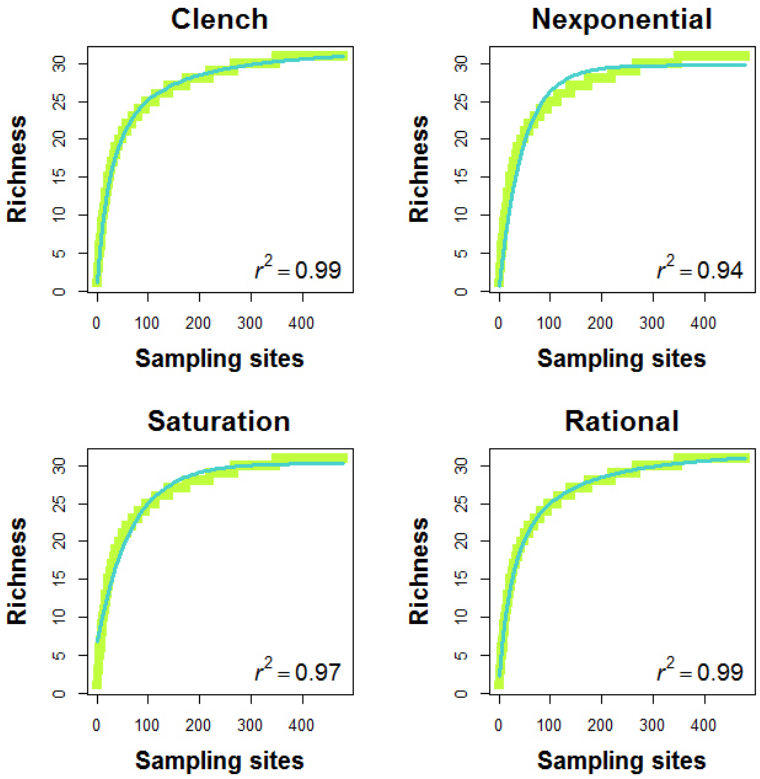
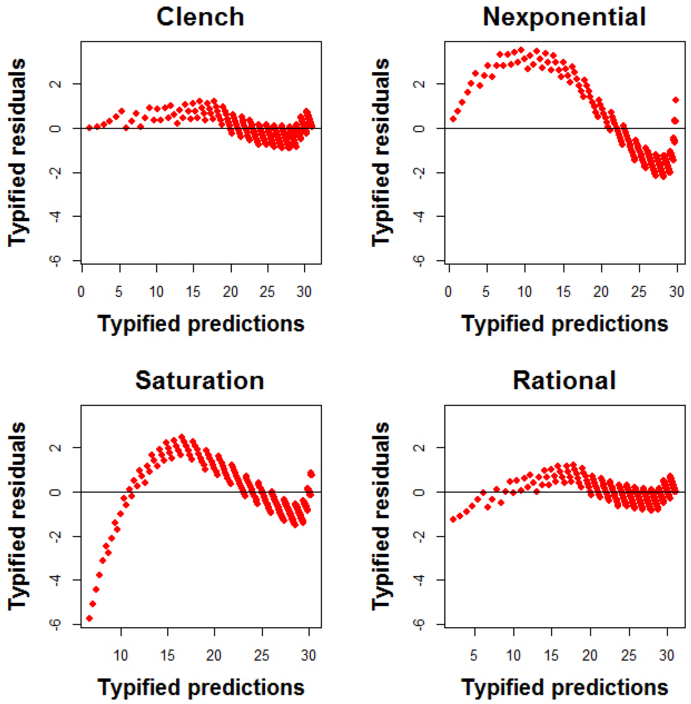
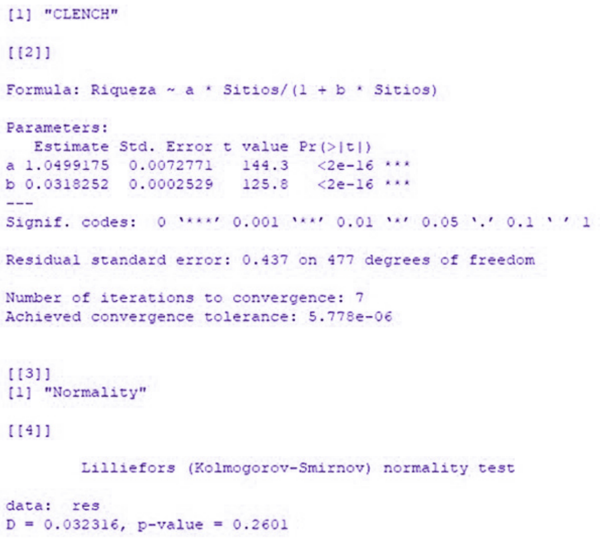


Figure XI2.6 shows that Clench and Rational functions also have the smallest residuals, have a nearly linear behavior and are the same throughout the entire range of predicted values for the dependent variable. Therefore, both functions are those that best fit.

Figure XI2.6. Relationship between the predicted values and the residuals typified from all accumulation functions.



The function of Clench’s test and the Kolmogorov-Smirnov normality with Lilliefors correction is displayed below. The residuals have a Normal distribution with p = 0.26.



```
[1] "CLENCH"

[[2]]

Formula: Riqueza ~ a * Sitios/(1 + b * Sitios)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 1.0499175  0.0072771   144.3   <2e-16 ***
b 0.0318252  0.0002529   125.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.437 on 477 degrees of freedom

Number of iterations to convergence: 7
Achieved convergence tolerance: 5.778e-06


[[3]]
[1] "Normality"

[[4]]

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.032316, p-value = 0.2601
```

The equation would be as shown below, in such a way that considering 500 sampled sites the theoretical total richness of the area would be of 31 species.

$$Richness = \frac{1.05 * Sites}{1 + 0.032 * Sites}$$

EXAMPLE 2

Step 1.

The study consisted of cultivate species of the genus *Alexandrium* at different concentrations of phosphate. The objective was to determine the minimum concentration of phosphate necessary to ensure growth and the maximum rate of growth of the population.

Figure XI2.7 shows the functions obtained, where it is noted that the saturation curve is the one that has a greater r^2. It is also noted that it was not possible to find an adjustment for the Rational function, which can happen for any function depending on the type of data.

Figure XI2.7. Relationship between the concentration of phosphate and the rate of daily growth of population of a species of the genus *Alexandrium*

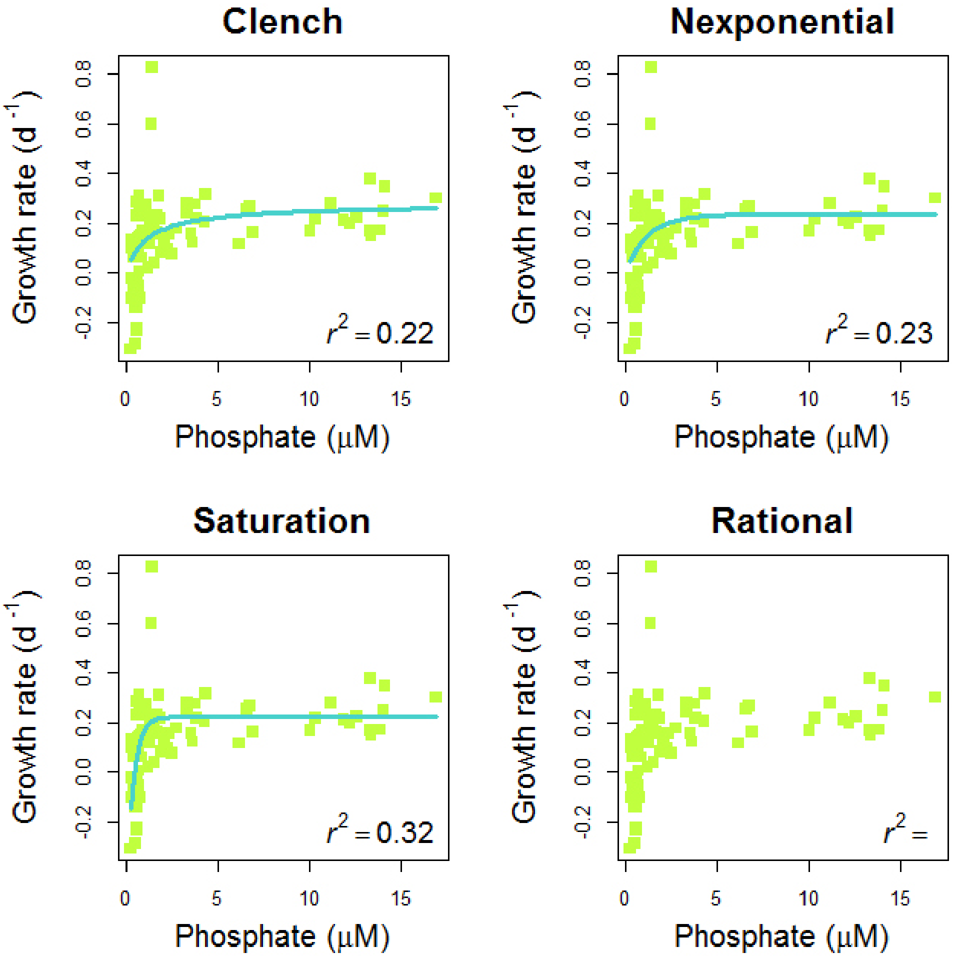
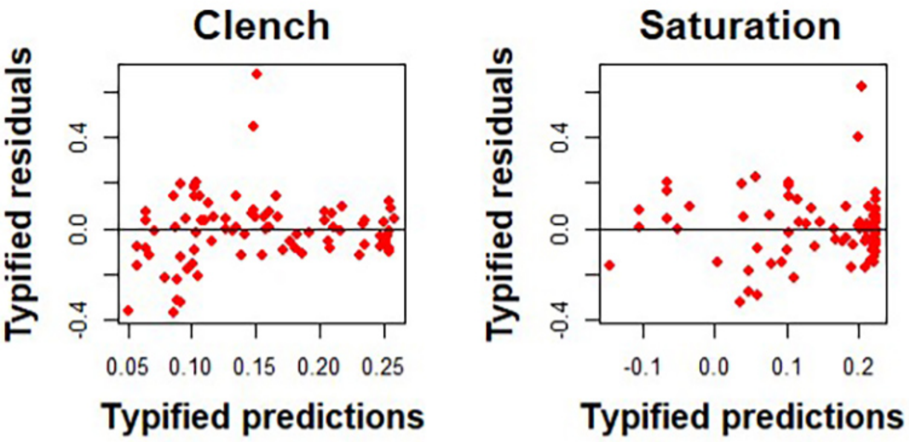


Figure XI2.8 shows that for the three functions, the residuals are small, have an almost linear behavior and are the same throughout the range of predicted values of the dependent variable. However, there appear to be outliers.

Figure XI2.8. Relationship between the predicted values and the residuals typified from all accumulation functions.



Step 2.

Figure XI2.8 it is noted that there were some outliers, which can be easily removed using the argument *outliers*. As it was previously observed the best model was the function of Saturation, specifying that outliers must be removed by using the argument *outliers="Saturación"*. Moreover, as it was not possible to adjust the Rational function, this was removed from the argument *model=c("Clench","Nexponential", "Saturation")*. Finally, as there are only three graphs, the argument *mfrow=c(1,3)* is used to represent them in one row and three columns.

Figure XI2.9 shows the obtained functions, where it is observed that the function of saturation is that which continues to have a greater r^2.

Figure XI2.9. Relationship between the concentration of phosphate and the rate of daily population growth of a species of the genus *Alexandrium*

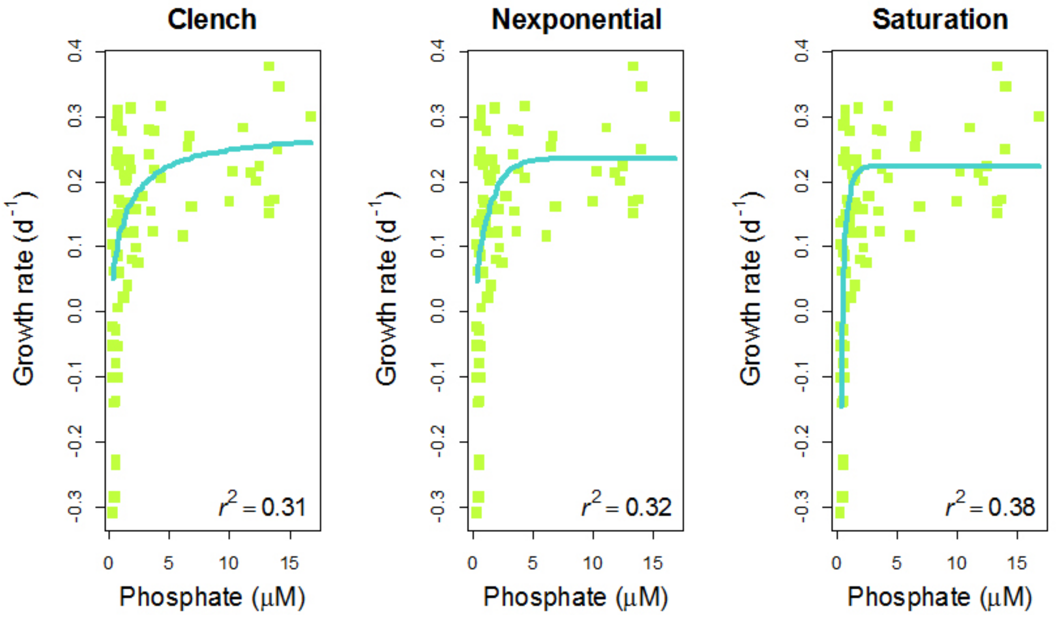
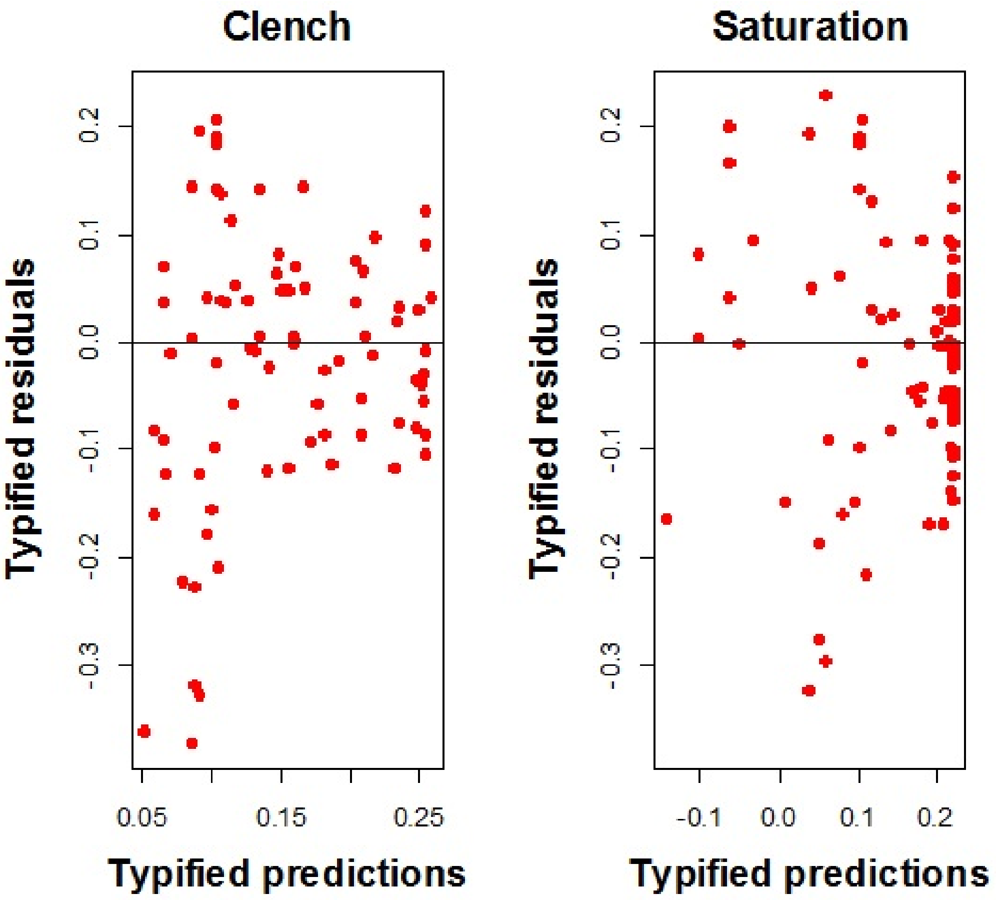
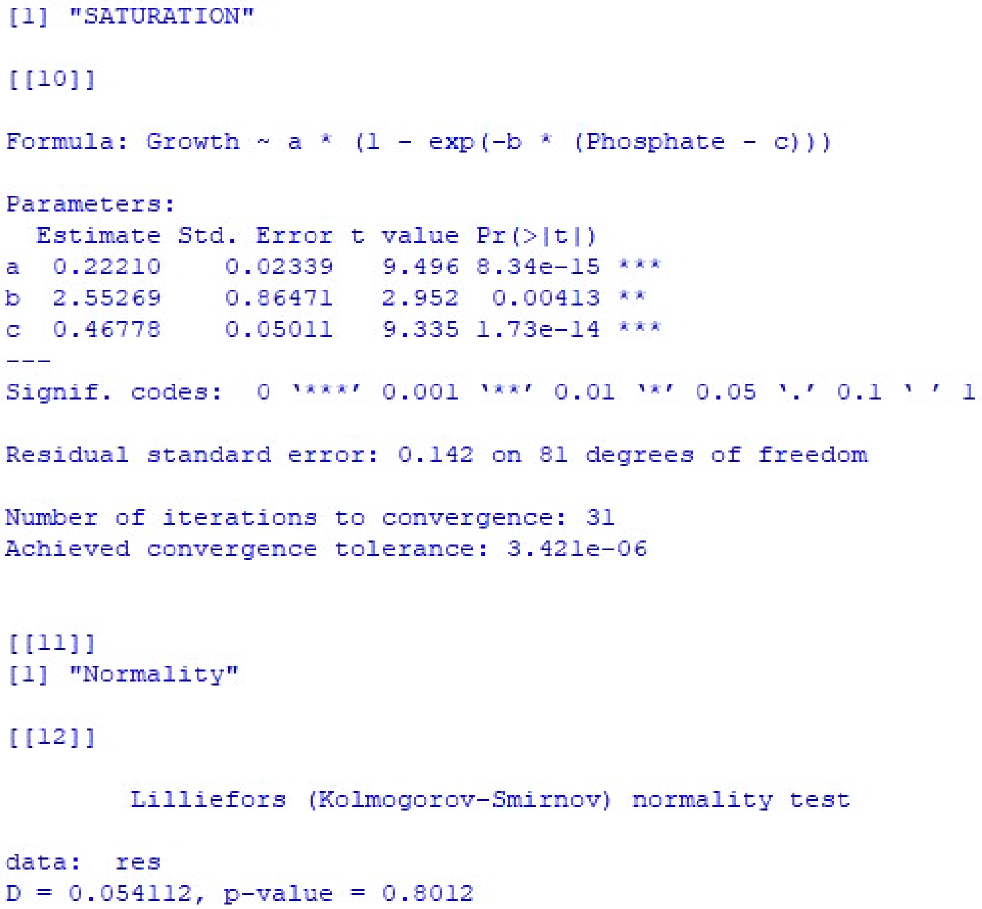


Figure XI2.10 shows that for the three functions, the residuals are homogeneous and, moreover, outliers were eliminated.

Figure XI2.10. Relationship between the predicted values and the residuals typified from all accumulation functions.



The function of Saturation, and the Normality test with the Kolmogorov-Smirnov Lilliefors' correction is shown below. The residuals have a Normal distribution with p = 0.801.



```
[1] "SATURATION"

[[10]]

Formula: Growth ~ a * (1 - exp(-b * (Phosphate - c)))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a  0.22210    0.02339   9.496 8.34e-15 ***
b  2.55269    0.86471   2.952  0.00413 **
c  0.46778    0.05011   9.335 1.73e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.142 on 81 degrees of freedom

Number of iterations to convergence: 31
Achieved convergence tolerance: 3.421e-06


[[11]]
[1] "Normality"

[[12]]

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.054112, p-value = 0.8012
```

The equation would be as shown below, so that the maximum growth rate is 0.222 d^{-1} and 0.468 μM is the minimum threshold of phosphate concentration required by the species to grow, as at this concentration of phosphate growth is zero.

$$y = 0.222\left(1 - e^{-2.553(Phosphate - 0.468)}\right)$$

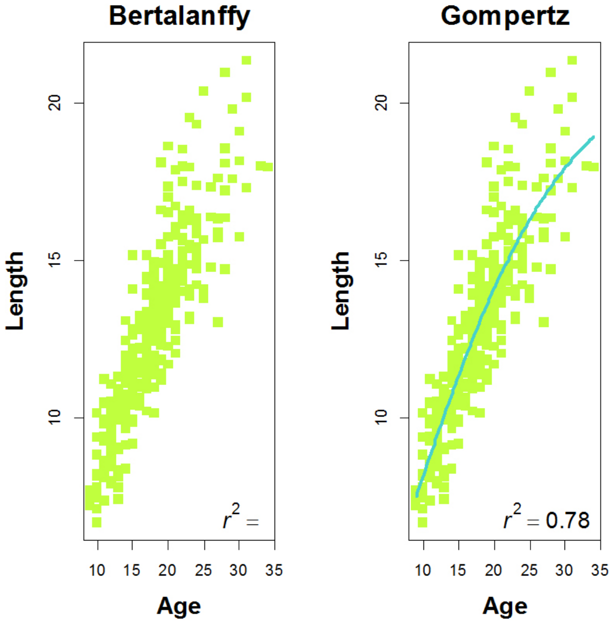
EXAMPLE 3

A study carried out with the anchovy (*Engraulis encrasicolus*), in the Alboran Sea and the Strait of Sicily, was to analyze the relationship between the length of the individuals and their age, which was quantified by counting otolith rings (Basilone et al., 2004). The aim was to compare the growth in both areas and determine the maximum length. For this reason, attempts had been made to find the growth function to best fit the data.

Step. 1

First the two growth functions with the argument are selected *model=c("Bertalanffy", "Gompertz")*. Figure XI2.11 shows that it could only find a setting with the Gompertz function.

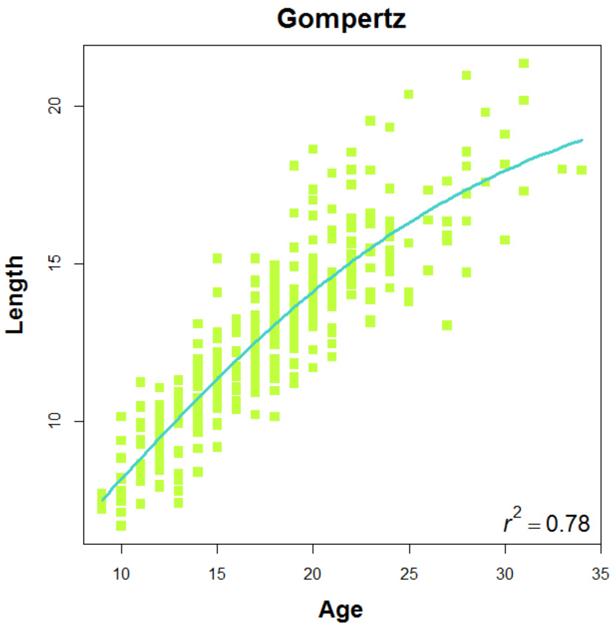
Figure XI2.11. Relationship between the length and age in the anchovy (*Engraulis encrasicolus*), in the Alboran Sea and the Strait of Sicily.



Step. 2

Then the model was made with only *model="Gompertz"* and to eliminate possible outliers using the argument *outliers="Gompertz"*. Figure XI2.12 displays the Gompertz function.

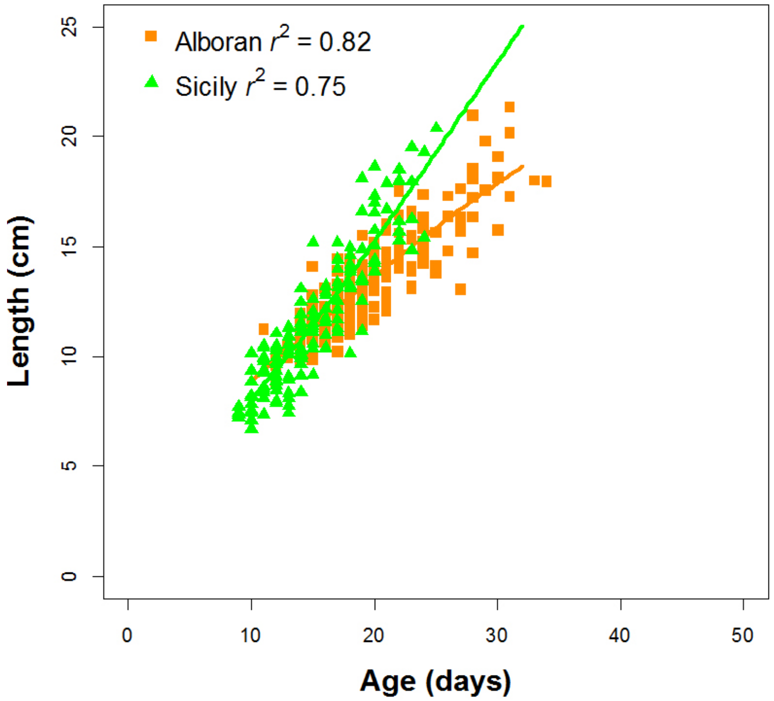
Figure XI2.12. Relationship between the length and age in the anchovy (*Engraulis encrasicolus*), in the Alboran Sea and the Strait of Sicily.



Step. 3

Finally, there was a script in order to show the trend in both areas. With the function *subset* the data of the Alboran Sea is first selected and the first graph is done, and with the second *subset* the data of Sicily is selected. With the function *seq*, a sequence of 10 to 32 with 100 divisions is generated, that allows then to obtain the values of *Length* by applying the Gompertz function corresponding to the data within each of the areas. With the function *lines* the trend line is represented. In order to have the coefficients of the functions of each of the areas, in step 2 there would be the script, but doing separately the Alborán Sea and the Strait of Sicily. Finally, with the function *legend* the legends are set. The resulting graph is shown in Figure XI2.13.

Figure XI2.13. Relationship between the length and age in the anchovy (*Engraulis encrasicolus*), in the Alboran Sea and the Strait of Sicily.



The equations obtained for each of the areas are as follows:

Area	Function
Alborán sea	$Length = 28.994 * e^{-1.8457 * e^{(-0.044718 * Age)}}$
Sicily	$Length = 57.613 * e^{-2.8974 * e^{(-0.038967 * Age)}}$

EXAMPLE 4

The growth of a population of phytoplankton normally follows a logistic growth curve. The purpose is to estimate the logistic curve of growth, to determine the maximum abundance that the species can reach under such growing conditions.

Figure XI2.14 shows that the adjustment to the logistics function is good, with a $r^2 = 0.98$, and residuals are homogeneous throughout the range of predicted values (Figure XI2.15).

Figure XI2.14. Evolution of the abundance of a species of phytoplankton over time

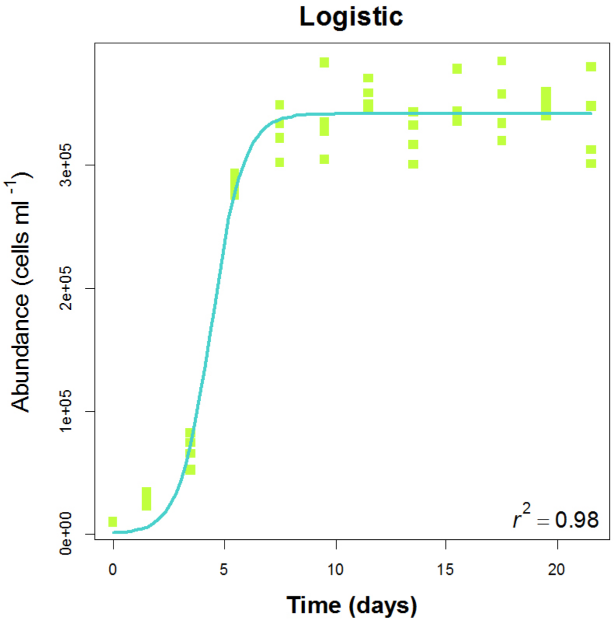
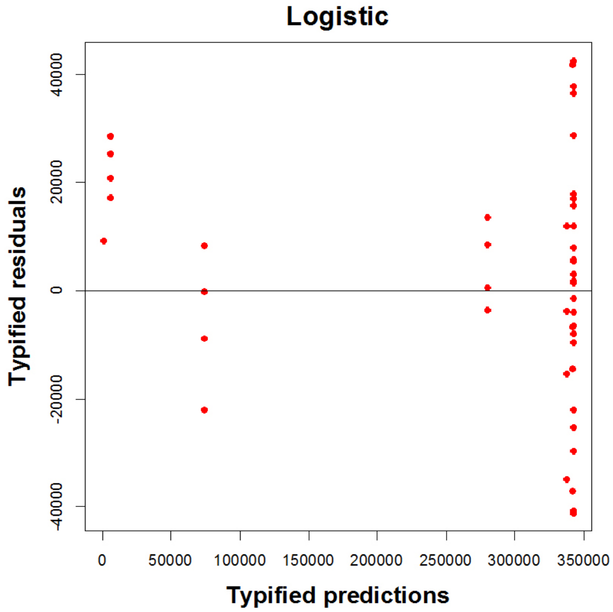
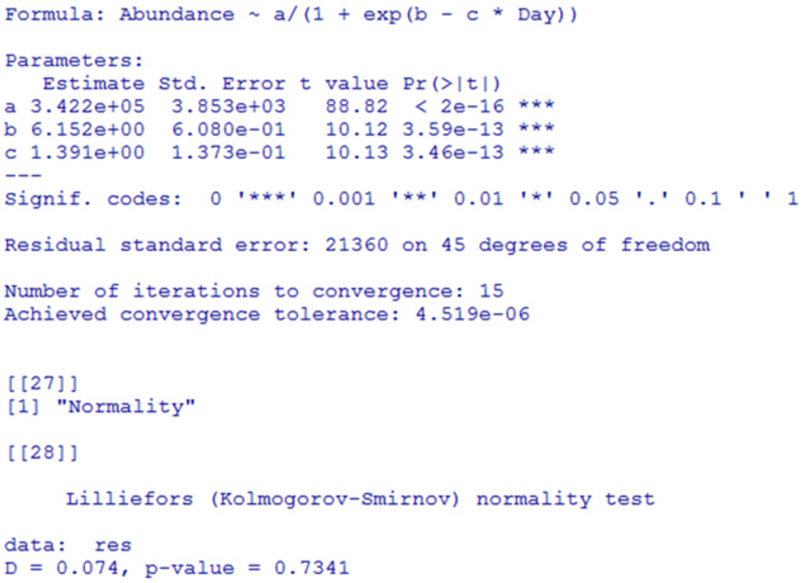


Figure XI2.15. Relationship between the predicted values and residuals typified in the function shown in Figure XI2.14.



The results show that the residuals are normal with p = 0.734 and logistic curve is defined by the following function, where it is observed that the maximum abundance is $3.422 * 10^5$ cells per ml:

$$Abundance = \frac{3,422 * 10^5}{1 + e^{(6,152 - 1,391 * Time)}}$$



```
Formula: Abundance ~ a/(1 + exp(b - c * Day))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 3.422e+05  3.853e+03   88.82  < 2e-16 ***
b 6.152e+00  6.080e-01   10.12 3.59e-13 ***
c 1.391e+00  1.373e-01   10.13 3.46e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21360 on 45 degrees of freedom

Number of iterations to convergence: 15
Achieved convergence tolerance: 4.519e-06


[[27]]
[1] "Normality"

[[28]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.074, p-value = 0.7341
```

Value

A TXT file is obtained with the results of the functions and the test of Normality for each of them, as well as graphics with the relationship between the dependent and independent variable in addition to graphics with the relationship between the predicted value and the residuals defined for each function.

References

Basilone, G., Guisande, C., Patti, B., Mazzola, S., Cuttitta, A., Bonanno, A. & Kallianiotis, A. (2004). Linking habitat conditions and growth in the European anchovy (*Engraulis encrasicolus*). *Fisheries Research*, 68: 9-19.

Callon, M., Courtial, J.P. & Penan, H. (1995) *Cienciometría: La medición de la actividad científica, de la bibliometría a la vigilancia tecnológica*. Trea, D.L., Gijón.

Clench, H.K. (1979) How to make regional lists of butterflies: some invoking empirically based criteria in selecting among thoughts. *The Journal of the Lepidopterists' Society*, 33: 216-231.

Flather, C.H. (1996) Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography*, 23: 155-168.

Frangópulos, M., Guisande, C., deBlas, E. y Maneiro, I. (2004) Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae*, 3: 131-139.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Hernández-Llamas, A. & Ratkowsky, D.A. (2004) Growth of fishes, crustaceans and molluscs: estimation of the von Bertalanffy, Logistic, Gompertz and Richards curves and a new growth model. *Marine Ecology Progress Series* 282: 237-244.

Miller, R.I. & Wiegert, R.G. (1989) Documenting completeness species-area relations, and the species-abundance distribution of a regional flora. *Ecology*, 70: 16-22.

Monod, J. (1950) La technique de culture continue, théorie et applications. *Annales de l'Institut Pasteur (Paris)*, 79: 390-410.

Ratkowski, D.A. (1990) *Handbook of nonlinear regression models*. Marcel Dekker, New York, 241 pp.

Smith, R.L. & Smith, T.M. (2000) *Ecología*. Pearson Educación, S.A., Madrid.

Examples

```
## Not run:

data(ZXI2)

#EXAMPLE 1. Accumulated richness and sampling sites

XI2(data=ZXI2, varX="Sampling.sites",varY="Richness", model=c("Clench",
"Nexponential", "Saturation","Rational"), mfrow=c(2,2), XLAB="Sampling sites")

#EXAMPLE 2. Growth rate of a toxic algae as a function
#of the concentration of phosphate

data(ZXI3)

#Step 1.
XI2(data=ZXI3, varX="Phosphate",varY="Growth", model=c("Clench", "Nexponential", "Saturation",
"Rational"), mfrow=c(2,2), YLAB=expression(paste("Growth rate (d "^"-1",")")),
XLAB=expression(paste("Phosphate (",mu,"M)")))

#Step 2.
XI2(data=ZXI3, varX="Phosphate",varY="Growth", model=c("Clench","Nexponential",
"Saturation"), outliers="Saturation",  mfrow=c(1,3),
YLAB=expression(paste("Growth rate (d "^"-1",")")),
XLAB=expression(paste("Phosphate (",mu,"M)")))

#EXAMPLE 3. Relationship between age and length in the anchovy

data(ZXI4)

#Step 1.
XI2(data=ZXI4, varX="Age",varY="Length",
model=c("Bertalanffy", "Gompertz"), mfrow=c(1,2))

#Step 2.

XI2(data=ZXI4, varX="Age",varY="Length",model="Gompertz",
outliers=c("Gompertz"), mfrow=NULL)

#Step 3

data(ZXI4)

#Alboran is selected
data1<-subset(ZXI4,ZXI4[,"Area"]
plot( formula = Length~Age, xlim = c(0,50), ylim = c(0,25),
xlab="Age (days)",
ylab="Length (cm)", font.lab=2, cex.lab=1.5, col = "#FF8C00FF" , pch = 15,
cex = 1.2, data = data1)

#Line of the function
cx<-seq(10,32, length.out = 100)
cy<-28.994*exp(-1.8457*exp(-0.044718*cx))
lines(cx,cy, col="#FF8C00FF",lwd=3 )

#Sicilia is selected
data2<-subset(ZXI4,ZXI4[,"Area"]
points(formula=Length~Age , col="#00FF00FF" , data=data2,
pch=17 , cex=1.2)

#Line of the function
cx<-seq(10,32, length.out = 100)
cy<-57.613*exp(-2.8974*exp(-0.038967*cx))
lines(cx,cy,  col="#00FF00FF",   lwd=3 )

#Legends
legend(x=0, y= 26, legend=expression(paste("Alboran ",
italic("r")^"2"," = 0.82")),pch=c(15), bty="n", cex=1.3,
col="#FF8C00FF")
legend(x=0, y= 24, legend=expression(paste("Sicily ",
italic("r")^"2"," = 0.75")),pch=c(17), bty="n", cex=1.3,
col="#00FF00FF")

#EXAMPLE 4. Evolution of the abundance of a population
#of phytoplankton species over time

data(ZXI5)

XI2(data=ZXI5, varX="Day",var="Abundance", model=c("Logistic"), mfrow=NULL,
YLAB=expression(paste("Abundance (cells ml"^"-1",")")),
XLAB="Time (days)")


## End(Not run)
```

XI3 *SELECTION OF NON-CORRELATED VARIABLES USING THE VIF*

Description

The inflation factor of the variance (VIF) is used to select uncorrelated variables, from an initial matrix with multiple variables.

Usage

```
XI3(data, variables, method="vifstep", threshold=10, varCode=NULL, file1="VIF.csv",
file2="Variables.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data` Data file.

`variables` Variables from which those non-correlated will be selected.

`method` There are three methods to select non-correlated variables: "vif", "vifcor" and "vifstep" (see section *details* for further information). "vif" method is not registered by default, despite being the recommended option, since it only works with RWizard and therefore does not work on other platforms, such as for example RGUI.

`threshold` Cut-off value for VIF.

`varCode` Optionally, variables of the original matrix can be selected, which are exported in the output file with non-correlated variables. For example, this allows to choose variables which are codes of rows.

`file1` CSV FILES. Name of the file with the VIF values for all variables.

`file2` CSV FILES. Name of the file where the non-correlated variables are exported.

`na` CSV FILES. Text that is used in the cells without data.

`dec` CSV FILES. Defines whether as decimal separator is used the comma "," or the dot ".".

`row.names` CSV FILES. Logical value which defines whether there are identifiers on the rows or a vector with a text for each of the rows.

Details

XI. REGRESSIONS

XI.1. REGRESSION MODELS FOR QUANTITATIVE DEPENDENT VARIABLES

XI.1.4. Variable selection using the inflation factor of variance (VIF)

XI.1.4.1. Selection of uncorrelated variables

In many multivariate statistical models is necessary to use variables that are not correlated. With this function, it is easy to generate a file with uncorrelated variables from a set of initial variables. The selection method is based on the inflation factor of variance (VIF).

A variable is redundant if is highly correlated with another or others, in which case it contains essentially the same information and should be excluded. The redundant variables in a model produced the effect of increasing the variance of the estimates of the coefficients (the coefficients are more unstable), thereby distorting the confidence intervals for these ratios and the contrasts of significance (may appear as a non-significant coefficient that it actually is). The problem is known as "multicollinearity," the regressors are interrelated and are not independent.

In general a certain degree of multicollinearity is inevitable, and should be accepted as such. Only when the multicollinearity is serious, action must be taken by removing from the model the variables that cause it.

The most widely used method to detect multicollinearity and redundant variables is the one that uses the so-called "inflation factor of variance" VIF: $VIF(x) = \frac{1}{(1-R^2)}$ being R^2 the coefficient of determination of the regression of the variable *x* with the remaining (square of the multiple correlation coefficient).

As $1 - R^2$ represents the fraction of variance of x not explained by the other variables, its inverse indicates the degree of redundant information: a VIF = 40 means that the information in this variable already contained in other variables is 40 times greater than the different information or new that this brings. The minimum value of VIF is 1, and has no maximum (can be arbitrarily large).

The following values are often used as reference values of VIF: VIF < 10 mild; 10 < VIF < 30 moderate; VIF > 30 serious. It is advisable to delete redundant by the variables with VIF > 30.

Delete variables can make the model biased or incomplete, it is generally recommended caution in doing so: it is preferable to maintain a variable that provides supporting information (at the expense of increasing the variability or instability) to supress it, because the effect could be more serious, especially if the sample is large, since the sample size corrects the problem making to reduce the variability of the estimates.

In this function between three different methods can be chosen to select the non-correlated variables.

If "vif" is selected, there is a pop-up window where it will report on the variables with VIF above the threshold indicated by the user in the argument *threshold*, and the user is the one who decides which variables will be eliminated. The advantage of the manual method is that the user decides which variable to delete from the correlated pairs.

In the other two options "vifcor" and "vifstep", an automatic selection of the variables to exclude is carried out, which is logically much faster and more convenient, but it is always preferable to use the manual option, so that the user always has control over when choosing which variable is selected between two correlated.

If "vifcor" is selected, it first looks for the pair of variables that has the maximum linear correlation, and eliminates the one which has the larger VIF, repeating the process until there is no variable with a high coefficient of correlation with another variable.

If "vifstep" is selected, it calculates the VIF for all the variables, and excludes that which has a greater VIF, always considering the threshold defined by the user in the argument *threshold*, repeating the process until no variable is correlated with another.

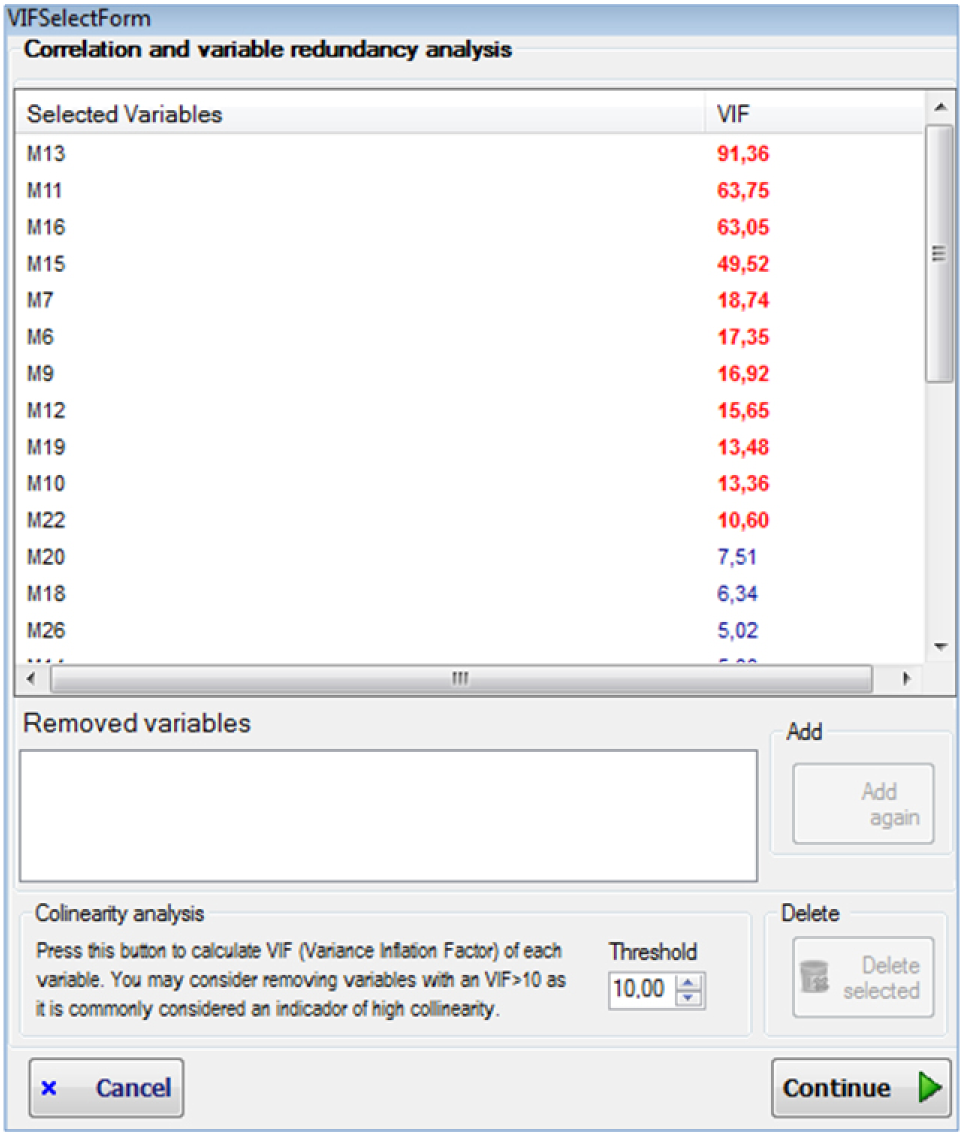
FUNCTIONS

The functions vif, vifcor and vifstep of the usdm package (Naimi, 2013; Naimi et al., 2014) were used.

EXAMPLE

Data are for morphometric measurements of fishes. The objective is to select those which are not correlated to a set of variables. The selection method chosen is "vif", which means it will get a menu which is shown below (Figure XI3.1), where the user is deciding which variables eliminate. Furthermore, the argument *varCode* indicates that the variables "Order", "Family", "Genus" and "Species" in the export file are included with the selection of uncorrelated variables

Figure XI3.1. Menu that allows to select which variables to eliminate, from those which are correlated.



VIFSelectForm

Correlation and variable redundancy analysis

Selected Variables	VIF
M13	91,36
M11	63,75
M16	63,05
M15	49,52
M7	18,74
M6	17,35
M9	16,92
M12	15,65
M19	13,48
M10	13,36
M22	10,60
M20	7,51
M18	6,34
M26	5,02

Colinearity analysis

Press this button to calculate VIF (Variance Inflation Factor) of each variable. You may consider removing variables with an VIF>10 as it is commonly considered an indicador of high collinearity.

Threshold 10,00

Value

Two CSV files are exported: one with VIF values of the variables that have been selected and are not correlated, and the other with the variables included in the *varCode* argument and those selected which are not correlated.

References

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. 2014. Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Examples

```
## Not run:

data(ZX3)
XI3(data=ZX3, variables = c("M2", "M3", "M4", "M5", "M6", "M7", "M8",
"M9", "M10", "M11", "M12", "M13", "M14", "M15", "M16", "M17", "M18",
"M19", "M20", "M21", "M22", "M23", "M24", "M25", "M26", "M27", "M28"),
method="vif",varCode=c("Order","Family","Genus","Species"))


## End(Not run)
```

XI4 *SELECTION OF CORRELATED VARIABLES USING VIF*

Description

The inflation factor of variance (VIF) is used to select highly correlated variables, from an initial array with multiple variables.

Usage

```
XI4(data, variables, threshold=10, varCode=NULL, file1="VIF.csv",
file2="Variables.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Data file.
`variables`	Variables from which those correlated will be selected.
`threshold`	Cut-off value for VIF.
`varCode`	Optionally, variables of the original matrix can be selected, which are exported in the output file with correlated variables. For example, this allows to choose variables which are codes of rows.
`file1`	CSV FILES. Name of the file with the VIF values for all variables.
`file2`	CSV FILES. Name of the file where the correlated variables are exported.
`na`	CSV FILES. Text that is used in the cells without data.
`dec`	CSV FILES. Defines whether as decimal separator is used the comma "," or the dot ".".
`row.names`	CSV FILES. Logical value which defines whether there are identifiers on the rows or a vector with a text for each of the rows.

Details

XI. REGRESSIONS

XI.1. REGRESSION MODELS FOR QUANTITATIVE DEPENDENT VARIABLES

XI.1.4. Variable selection using the inflation factor of variance (VIF)

XI.1.4.2. Selection of correlated variables

It is not frequent, but there are statistical multivariate models where it is useful to note that the variables are correlated, for example, the Principal Components Analysis (PCA). With this function it is easy to generate a file with the correlated variables from a set of initial variables. The method of selection is based on the inflation factor of the variance (VIF) that is explained in the section *details* of the function XI3.

FUNCTIONS

The function vif of the usdm package (Naimi, 2013; Naimi et al., 2014) was used.

EXAMPLE

Data correspond to morphometric measures of fishes. The goal is to select those variables that are highly correlated to a set of variables. In addition, it indicates that the variables "Family" and "Gender" be included in the export file with the selection of correlated variables.

Value

Two CSV files are exported: one with VIF values of the variables that have been selected and are correlated, and the other with the variables included in the *varCode* argument and those selected which are correlated.

References

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Examples

```
## Not run:

data(ZX3)

XI4(data=ZX3, variables=c("M11","M12","M13","M15","M24","M2", "M3", "M4", "M5"),
varCode=c("Family","Genus"))


## End(Not run)
```

XI5 *MULTIPLE REGRESSION*

Description

Different statistical techniques are applied to obtain the type of function that best fits the relationship that exists between a quantitative dependent variable and multiple quantitative independent variables.

Usage

```
XI5(data, varY, varX, outliers=FALSE, quant1=0.05, quant2 = 0.95,
stepwise=FALSE,
direction="both", deplog=FALSE, indlog=FALSE, method="hier.part",
threshold=10,
CoVa=FALSE, typeV="lmg", rela=TRUE, b=1000, names.abbrev=15, ylimV=NULL,
mainV=NULL, cex.title=1.5, PAIRS=TRUE, lower.panel=panel.smooth,
upper.panel=panel.reg,
diag.panel=panel.hist, main=NULL, cex.main=2, cex=1.2, pch=24, colhist="#00FFFFFF",
bg="#B2DFEEFF", cex.labels=1, font.labels=2, file1="Output.txt", file2="Residuals.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data` Data file.

`varY` Dependent variable.

`varX` Independent variables.

`outliers` If it is TRUE, the outliers are removed using the selected regression model (for more details see section *details* of function XI1).

`quant1` Quantile of the lower end for the elimination of the atypical data.

`quant2` Quantile of the top end for the elimination of atypical data.

`stepwise` If TRUE, a regression to eliminate those variables that are not significant is therefore applied by steps. It uses Akaike Information Criterion (AIC) to define which are the variables that are excluded (see section *details* for more information).

`direction` The mode of stepwise search, can be one of the following: "both" (default), "backward", or "forward"

`deplog` If TRUE, the logarithm is applied to the dependent variable by using the equation $ln(variable - (min(variable) - 1equation))$.

`indlog` If TRUE, the logarithm of all independent variables is calculated and added to the rest of variables, to perform the regression model with transformed and untransformed independent variables, in order to get the best model. The same equation used in the argument *deplog* is applied.

`method` There are four methods to select the non-correlated variables: "vif", "vifcor" "vifstep" and "hier.part" (default). The "vif" method is not selected by default, because it can only work with RWizard and, therefore, it does not work in other platforms such as RGUI. If this argument is NULL, the variables that show collinearity are not deleted. For more information, see section *details*.

`threshold` Cut-off value for the VIF (for more information, see section *details* of the function XI3).

`CoVa` If it is TRUE, the graph is shown with the contribution of independent variables to the regression model, provided that the number of independent variables is greater than two.

`typeV` GRAPHIC CONTRIBUTION OF VARIABLES OF ARGUMENT *CoVa*. It can be one or a vector with the following methods:"lmg", "pmvd" (only version outside the United States), "last", "first", "betasq", "pratt", "genizi" or "car". For more details see section *details* of the function calc.relimp.

`rela` GRAPHIC CONTRIBUTION OF VARIABLES OF ARGUMENT *CoVa*. If TRUE the relative importance of all variables add up to 100%.

`b` GRAPHIC CONTRIBUTION OF VARIABLES OF ARGUMENT *CoVa*. It is the number of boot cycles requested at boot.relimp. Be sure to adjust to a lower number, if you are simply testing code.

`names.abbrev` GRAPHIC CONTRIBUTION OF VARIABLES OF ARGUMENT *CoVa*. An integer which defines the number of letters in the labels.

`ylimV` GRAPHIC CONTRIBUTION OF VARIABLES OF ARGUMENT *CoVa*. A vector with Y axis limits.

`mainV` GRAPHIC CONTRIBUTION OF VARIABLES OF ARGUMENT *CoVa*. Title of the graph.

`cex.title` GRAPHIC CONTRIBUTION OF VARIABLES OF ARGUMENT *CoVa*. Text size title.

`PAIRS` If TRUE, the matrix chart is displayed. If the number of variables is greater than 10, it is better to define the argument as FALSE, because the utility is lost when the relationships among the variables are not well observed.

`lower.panel` MATRIX CHART. Graph type of the upper left panel.

`upper.panel` MATRIX CHART. Graph type of the upper right panel.

`diag.panel` MATRIX CHART. The diagonal graph type.

`main` MATRIX CHART. Graph title.

`cex.main` MATRIX CHART. Size of the letter of the title.

`cex` MATRIX CHART. Size of the symbols.

`pch` MATRIX CHART. Type of symbol.

`colhist` MATRIX CHART. Color of the bars of the histogram.

`bg` MATRIX CHART. Color of symbols.

`cex.labels` MATRIX CHART. Size of the labels.

`font.labels` MATRIX CHART. Label font type.

`file1` TXT FILE. Name of the output file with the results of the analysis.

`file2`	CSV FILE. Name of the file with the residuals of the regression models.
`na`	CSV FILES. Text used in cells without data.
`dec`	CSV FILES. It defines if a comma "," or a dot "." is used as decimal separator.
`row.names`	CSV FILES. Logical value which defines whether there are identifiers on the rows or a vector with a text for each of the rows.

Details

XI. REGRESSIONS

XI.1. REGRESSION MODELS FOR QUANTITATIVE DEPENDENT VARIABLES

XI.1.5. Multiple regression

So far we have seen regressions in which there was a single independent variable. However, it is often necessary to consider whether our dependent variable is related to more than one variable. In this case, the multiple regression should be used.

In addition to normal and homoscedasticity of the residuals, and that there should be no autocorrelation, requirements of simple regressions (see function XI1), another requirement of the multiple regressions is that there should be no correlation between independent variables, i.e., there should be no multicollinearity.

In the function four different methods for selecting uncorrelated variables can be chosen. If "vif" is selected, a pop-up window comes out where variables with VIF above threshold indicated by the user in the argument *threshold* are reported, and the user decides which variables are removed. The advantage of the manual method is that the user decides which variables to delete from the correlated pairs.

In the "vifcor" and "vifstep" options, an automatic selection of the variables to exclude is made, which is obviously much faster and comfortable. If "vifcor" is selected, you must first find the pair of variables that has the maximum linear correlation, and eliminate those having the greater VIF, and the process should be repeated until there is no variable with a high correlation coefficient with another variable. If "vifstep" is selected, the VIF must be calculated for all variables, and exclude which has the greater VIF, always considering the threshold defined by the user in the argument *threshold*, the process should be repeated until no variable is correlated with another. The problem of "vifcor" and "vifstep" methods is that these are not taken into account when eliminating variables, which is the one that has a greater contribution to explain the observed variability in the dependent variable and, therefore, are not the best methods for obtaining the best regression model.

The default is "hier.part" in which the contribution of all independent variables to the regression model is calculated by the hierarchical partitioning method (Chevan & Sutherland, 1991) and subsequently between the variables that are above VIF threshold, eliminating those that contribute less to the regression model. Therefore, among the auto-correlated variables, those that best explain the observed variability in the dependent variable are selected. The problem is that one can only use this method if the number of dependent variables, including those log-transformed and untransformed, is not greater than 12. If the number is greater, then the "vifstep" method is automatically applied.

The default option of the argument *outliers=TRUE*, means that outliers are eliminated as explained in function XI1.

If the argument *stepwise=TRUE* is selected, it means that a regression is performed by steps using Akaike Information Criterion (*AIC*) to define which are the variables that are excluded from the regression model, since they do not contribute to explain the dependent variable. The criterion of

Akaike (*AIC*) is used to decide how many explanatory variables are chosen, and what they should be. The method of Akaike values the goodness of the model by setting a penalty because of their complexity, so a simpler model is preferable to another with more independent variables that only explains a small additional portion of the variability of the dependent variable. In general $AIC = 2k - 2ln(L)$, where *k* is the number of explanatory variables or parameters of the model and *L* the likelihood or probability associated with the sample used in accordance with model, so that *AIC* is smaller than the lower the number of variables and the greater the likelihood; between alternative models the one with the lowest value of *AIC* must be chosen. At each step the variable that, when it is incorporated into the model, makes smaller the value *AIC* is chosen, until any unused variables allows to reduce it.

If the argument *indlog* = *TRUE*, the logarithm of all independent variables were calculated using the following equation $LnVariable = ln(variable - (min(variable) - 1))$, and all these variables are added to the non-transformed ones. The aim is to perform a regression model in which the transformed and non-transformed variables are included, to find the best model, because sometimes the transformed variable shows a better relationship with the dependent variable that the not transformed logarithmically.

FUNCTIONS

The hierarchical partitioning is performed with the function hier.part of the hier.part package (Walsh & Mac Nally, 2014)). The lillie.test function of the nortest package (Gross, 2013) is used to perform the Normality test of Kolmogorov-Smirnov with Lilliefors' correction, the function dwtest of the lmtest package (Hothorn et al., 2013) to analyze the autocorrelation with the test of Durbin-Watson statistic, the function bptest of package lmtest (Hothorn et al. , 2013) to perform the Breusch-Pagan test of homoscedasticity, functions vif, vifcor and vifstep of the package usdm (Naimi, 2013; Naimi et al., 2014) for the analysis of the VIF, the function calc.relimp for the graph of the contribution of the independent variables (Grömping, 2006; 2013) and the functions panels and panels.diag for the matrix chart (Grosjean, 2013).

EXAMPLE

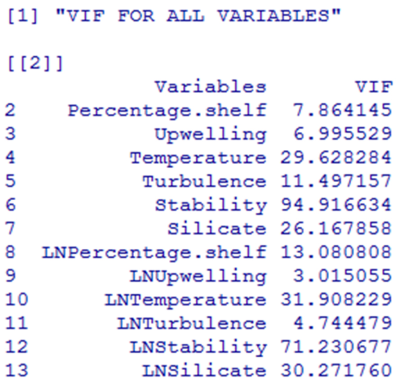
The data correspond to a study in areas of upwelling where information on the concentration of chlorophyll (mg^{-3}) was obtained, percentage of continental shelf in the cell, transport of Ekman ($m^3km^{-1}s^{-1}$), which is a measure of upwelling, temperature, turbulence (m^3s^{-1}), stability of the water column ($cycless^{-1}$) and concentrations of phosphate, nitrate and silicate in μMl^{-1}. The objective is to determine which are the variables that best explain the observed variations in the concentration of chlorophyll.

Step 1.

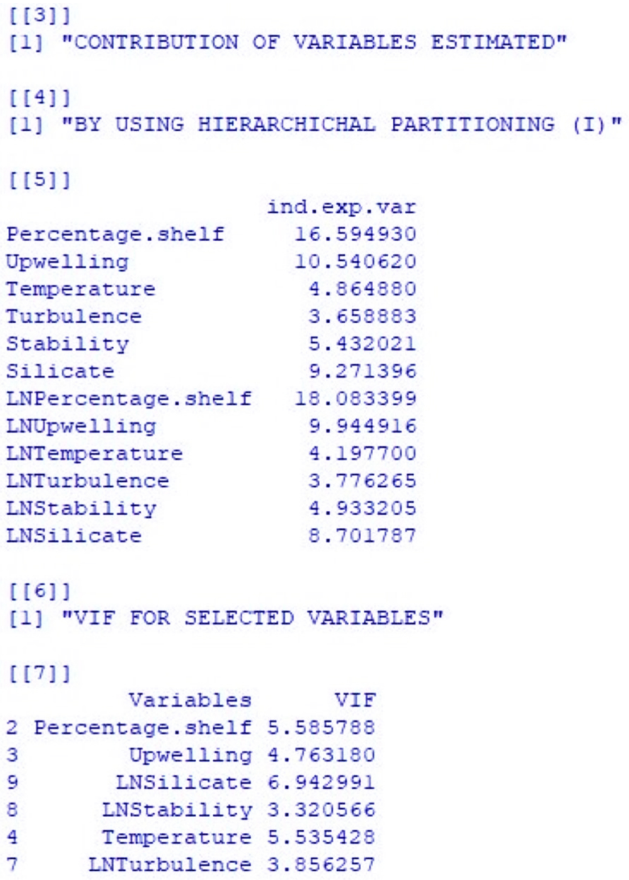
First working with non-transformed dependent variable, and therefore leaving the option that comes by default in the argument *deplog=FALSE*. However, the argument *outliers=TRUE* to remove atypical data, *stepwise=TRUE* to perform regression by steps and delete non-significant variables model, *indlog=TRUE* to apply the logarithm of the independent variables and add them to the other variables to estimate the model, and leave the default option *method="hier.part"* to remove the correlated independent variables using the hierarchical partitioning methods.

The first table shows high values of VIF, which means that there is multicollinearity between the independent variables, which is logical since they are the same transformed and non-transformed variable, for example «LNTemperature» and «Temperature».

The second table shows the contribution of each of the variables to the regression model estimated by the hierarchical partitioning method. As in the script, the default option *method="hier.part"* is used, the auto-correlated variables (variables with VIF values greater than the threshold) are removed following the criteria to eliminate those with less contribution to the model between the auto-correlated. The third table shows the variables that have been selected, once eliminated the auto-correlated of minor contribution.



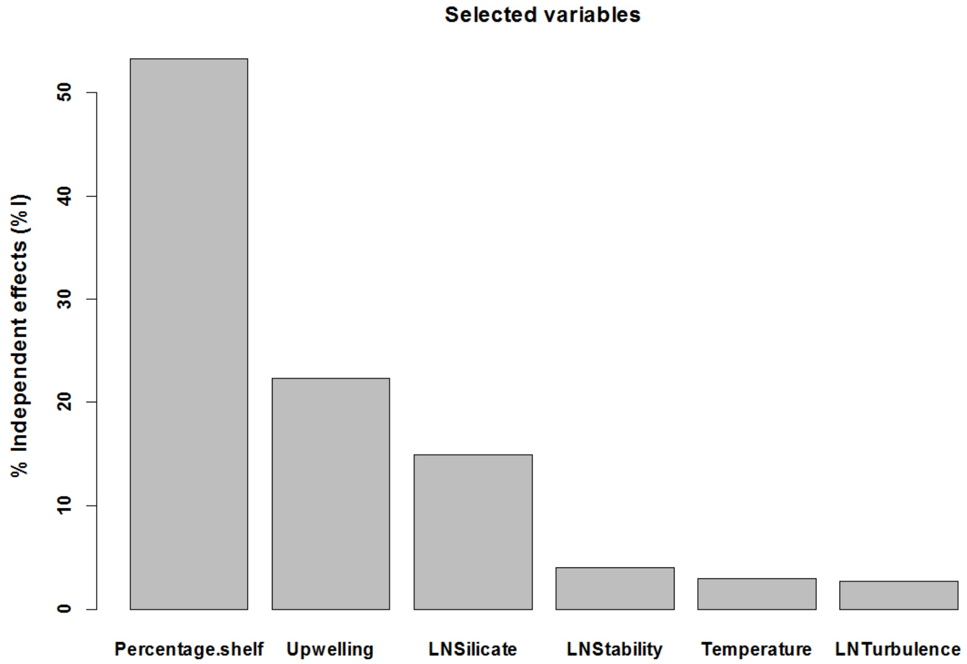
```
[1] "VIF FOR ALL VARIABLES"

[[2]]
             Variables       VIF
2     Percentage.shelf  7.864145
3            Upwelling  6.995529
4          Temperature 29.628284
5           Turbulence 11.497157
6            Stability 94.916634
7             Silicate 26.167858
8   LNPercentage.shelf 13.080808
9          LNUpwelling  3.015055
10       LNTemperature 31.908229
11        LNTurbulence  4.744479
12         LNStability 71.230677
13          LNSilicate 30.271760
```



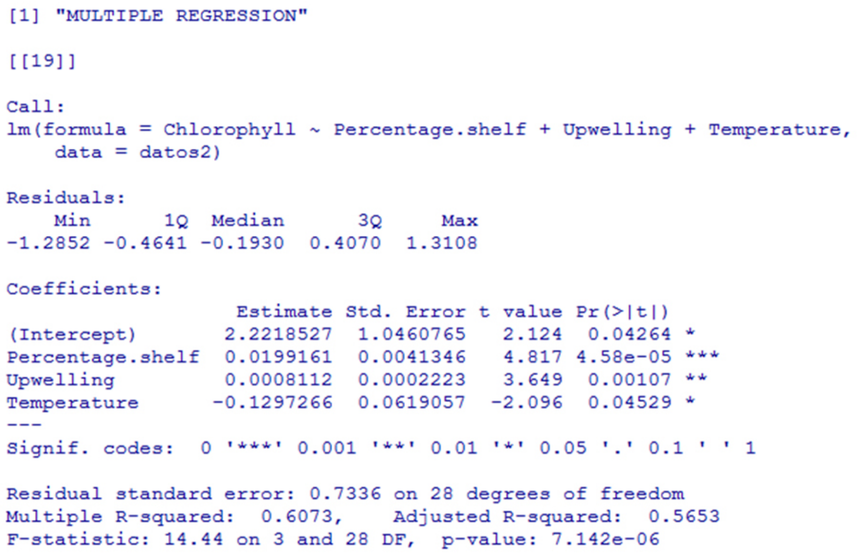
```
[[3]]
[1] "CONTRIBUTION OF VARIABLES ESTIMATED"

[[4]]
[1] "BY USING HIERARCHICHAL PARTITIONING (I)"

[[5]]
                   ind.exp.var
Percentage.shelf     16.594930
Upwelling            10.540620
Temperature           4.864880
Turbulence            3.658883
Stability             5.432021
Silicate              9.271396
LNPercentage.shelf   18.083399
LNUpwelling           9.944916
LNTemperature         4.197700
LNTurbulence          3.776265
LNStability           4.933205
LNSilicate            8.701787

[[6]]
[1] "VIF FOR SELECTED VARIABLES"

[[7]]
         Variables      VIF
2 Percentage.shelf 5.585788
3        Upwelling 4.763180
9       LNSilicate 6.942991
8      LNStability 3.320566
4      Temperature 5.535428
7     LNTurbulence 3.856257
```

Figure XI5.1 shows the contribution of each of the selected variables, estimated by the hierarchical partitioning method. The percentage of continental shelf in the cell was the variable that most contributes to explain the observed variability in the concentration of chlorophyll and secondly the intensity of upwelling.

Figure XI5.1. Relative contribution of independent variables in the regression model.

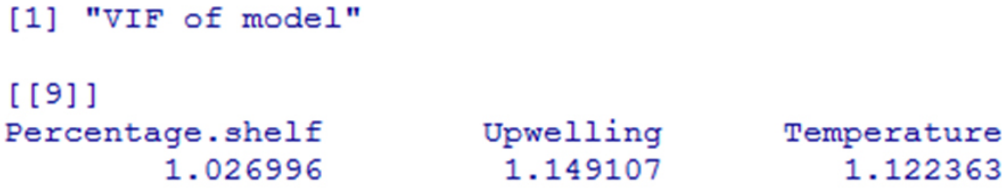


The variables selected and shown in Figure XI5.1, not all were significant and in the stepwise regression model only the percentage of continental shelf that exists in the cell, the intensity of upwelling and temperature were selected. These three variables explained 60.7% of the observed variance in chlorophyll concentration ($r^2 = 0.61$).



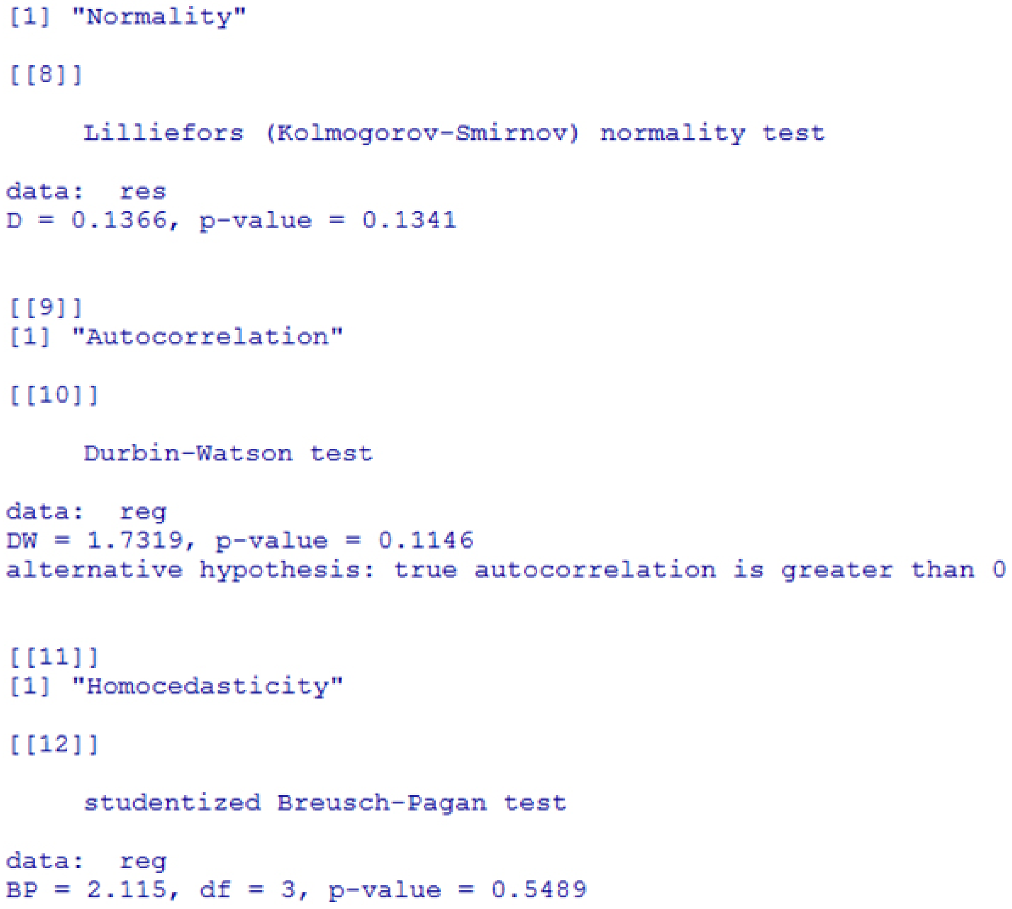
```
[1] "MULTIPLE REGRESSION"

[[19]]

Call:
lm(formula = Chlorophyll ~ Percentage.shelf + Upwelling + Temperature,
    data = datos2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2852 -0.4641 -0.1930  0.4070  1.3108 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.2218527  1.0460765   2.124  0.04264 *  
Percentage.shelf  0.0199161  0.0041346   4.817 4.58e-05 ***
Upwelling         0.0008112  0.0002223   3.649  0.00107 ** 
Temperature      -0.1297266  0.0619057  -2.096  0.04529 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7336 on 28 degrees of freedom
Multiple R-squared:  0.6073,	Adjusted R-squared:  0.5653 
F-statistic: 14.44 on 3 and 28 DF,  p-value: 7.142e-06
```

The three selected variables have VIF values smaller than 10 and, therefore, there is no problem with the collinearity.



```
[1] "VIF of model"

[[9]]
Percentage.shelf        Upwelling      Temperature 
        1.026996         1.149107         1.122363 
```

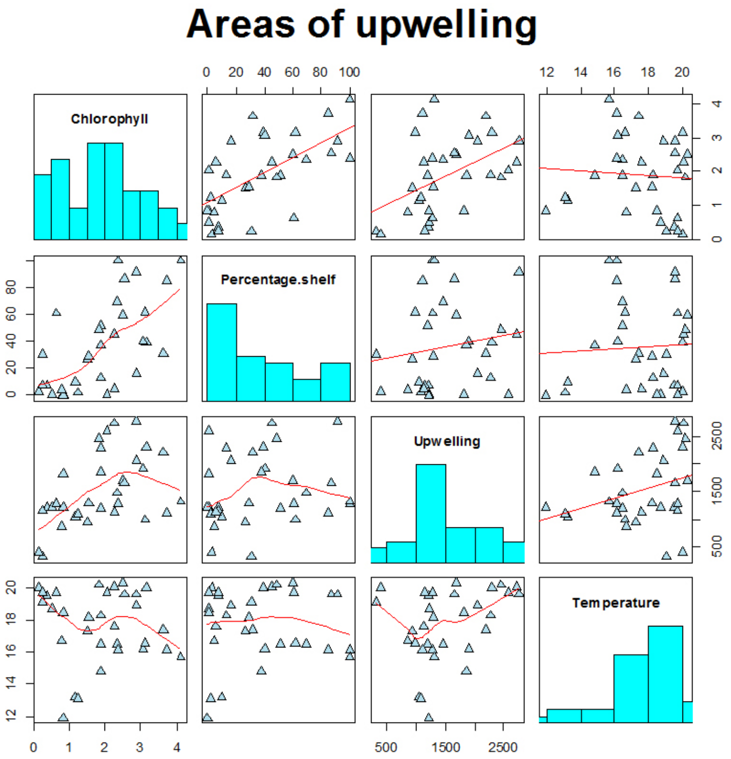
It is also noted, that the assumptions of normality (p = 0.134) are fulfilled, there is no autocorrelation (p = 0.115) and there are homoscedasticity in residuals (p = 0.549). In case of non compliance with some of these assumptions see section *details* of the function XI1, for more information.



```
[1] "Normality"

[[8]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.1366, p-value = 0.1341


[[9]]
[1] "Autocorrelation"

[[10]]

     Durbin-Watson test

data:  reg
DW = 1.7319, p-value = 0.1146
alternative hypothesis: true autocorrelation is greater than 0


[[11]]
[1] "Homocedasticity"

[[12]]

     studentized Breusch-Pagan test

data:  reg
BP = 2.115, df = 3, p-value = 0.5489
```

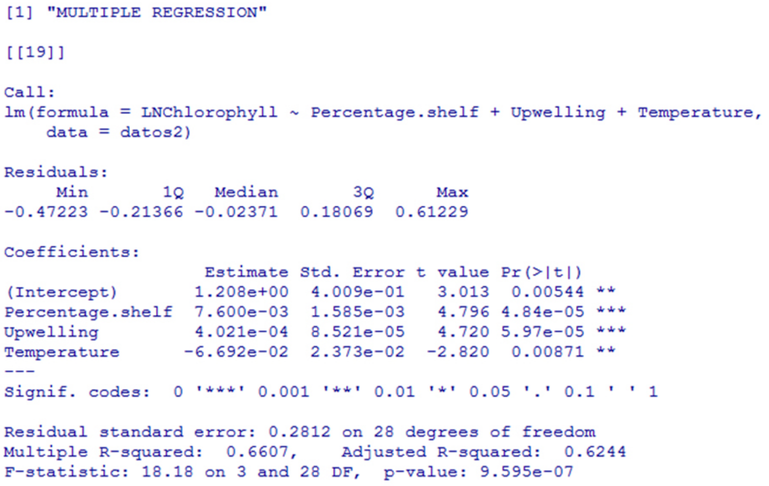
Figure XI5.2 shows the relationship that exists between all the selected variables, including the chlorophyll and the independent variables.

Figure XI5.2. Matrix graph that shows the relationship between theindependent variables and chlorophyll concentration.



Step 2.

We then repeat the process but changing the dependent variable and, to this end, we modify the argument *deplog=TRUE*. The regression model obtained shows that the same variables are chosen as they were previously selected: the percentage of shelf and upwelling are positively related and temperature negatively with the chlorophyll concentration.

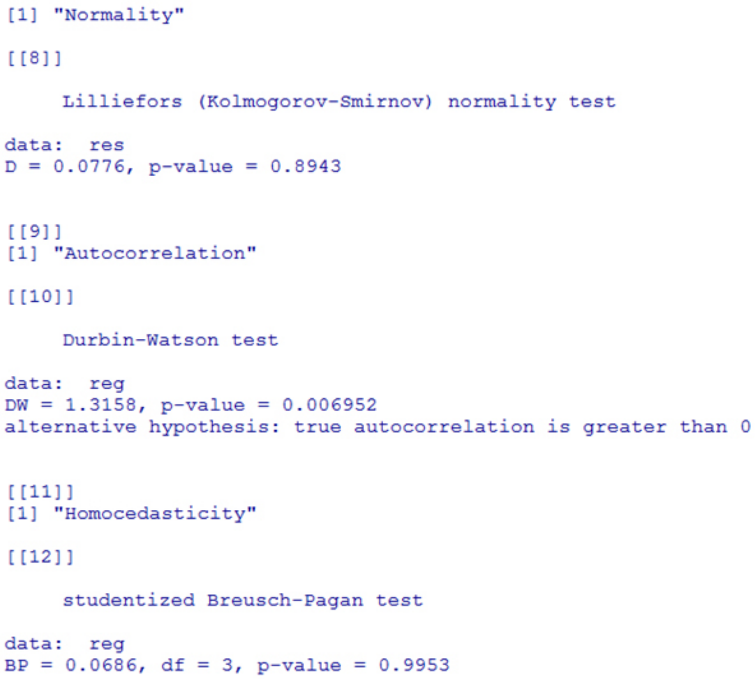


```
[1] "MULTIPLE REGRESSION"

[[19]]

Call:
lm(formula = LNChlorophyll ~ Percentage.shelf + Upwelling + Temperature,
    data = datos2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47223 -0.21366 -0.02371  0.18069  0.61229

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.208e+00  4.009e-01   3.013  0.00544 **
Percentage.shelf  7.600e-03  1.585e-03   4.796 4.84e-05 ***
Upwelling         4.021e-04  8.521e-05   4.720 5.97e-05 ***
Temperature      -6.692e-02  2.373e-02  -2.820  0.00871 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2812 on 28 degrees of freedom
Multiple R-squared:  0.6607,    Adjusted R-squared:  0.6244
F-statistic: 18.18 on 3 and 28 DF,  p-value: 9.595e-07
```

The selected variables explain a 66% of the variability observed in the concentration of chlorophyll

($r^2 = 0.66$). Therefore, this model is better than the last, in which chlorophyll variable was not transformed. It is possible that explanatory variables are included in the model with p > 0.05. With the Akaike criterion used, this means that the greater complexity of the model by adding this variable is compensated by the increase in likelihood (the observed sample is most likely or credible with the new model). However, it should be noted, that other methods of selection step by step discarded all the non-significant variables, so that the researcher's decision to exclude a variable with p > 0.05 could also be considered reasonable.

It is also noted, that the normality assumptions are fulfilled (p = 0.894), there is a certain autocorrelation (p = 0.007, see section *details* function XI1 for details) and there are homoscedasticity in residuals (p = 0.995).



```
[1] "Normality"

[[8]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  res
D = 0.0776, p-value = 0.8943


[[9]]
[1] "Autocorrelation"

[[10]]

     Durbin-Watson test

data:  reg
DW = 1.3158, p-value = 0.006952
alternative hypothesis: true autocorrelation is greater than 0


[[11]]
[1] "Homocedasticity"

[[12]]

     studentized Breusch-Pagan test

data:  reg
BP = 0.0686, df = 3, p-value = 0.9953
```

The regression model is defined by the following formula:

$$ln(Chlorophyll - (min(Chlorohyll) - 1)) = 1.2 + 0.007 * Percentage.shelf +$$

$$0.0004 * Upwelling - 0.069 * Temperature$$

Value

A TXT file is obtained with the VIF values of all variables, the VIF of the selected variables after removing those with colinearity, the VIF of the variables that are finally in the regression model, the normality test, homogeneity of variances and homoscedasticity and the regression model. A matrix of scatterplots, a plot with the relative contribution of the independent variables estimated by using hierarchical partitioning and a plot with the relative contribution of the independent variables in the multiple regression if the argument *CoVa=TRUE*. A CSV file is also obtained with the standardized residuals.

References

Chevan, A. & Sutherland, M. (1991) Hierarchical Partitioning. *The American Statistician* 45, 90-96.

Durbin, J. & Watson G.S. (1951) Testing for serial correlation in least squares regression. *Biometrika*, 38: 159-171.

Grömping, U. (2006) Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, 17: 1-27.

Grömping, U. (2013)Relative importance of regressors in linear models. R package version 2.2-2. Available at: http://CRAN.R-project.org/package=relaimpo.

Grosjean, P. (2013) SciViews GUI API - Main package. R package version 0.9-5. Available at: http://CRAN.R-project.org/package=SciViews.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Hothorn, T. et al., (2013) Testing Linear Regression Models R package version 0.9-33. Available at: http://CRAN.R-project.org/package=lmtest.

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Walsh, C. & Mac Nally, R. (2014) Hierarchical Partitioning. R package version 1.0-4. Available at: http://CRAN.R-project.org/package=hier.part.

Examples

```
## Not run:
data(ZXI6)

#Step 1

XI5(data=ZXI6, varY="Chlorophyll", varX = c("Percentage.shelf",
"Upwelling", "Temperature", "Turbulence", "Stability", "Silicate"),
outliers=TRUE, stepwise=TRUE, indlog=TRUE, main="Areas of upwelling")

#Step 2

XI5(data=ZXI6, varY="Chlorophyll",varX = c("Percentage.shelf",
"Upwelling",
"Temperature", "Turbulence", "Stability", "Silicate"), outliers=TRUE,
stepwise=TRUE, deplog=TRUE, indlog=TRUE, main="Areas of upwelling")

## End(Not run)
```

XI6 *BINOMIAL LOGISTIC REGRESSION-ESTIMATION*

Description

Apply various statistical techniques to obtain the type of function that best fits the possible relationship between a qualitative dependent variable with two categories and several independent qualitative or quantitative.

Usage

```
XI6(data, cat, var, resp, stepwise=TRUE, cart=TRUE, minsplit=2, minbucket=2,
cp=0.0001, surrogatestyle=0, type=0, extra=2, varlen=0, ycompress= TRUE,
main=NULL, cex.main=2, cex=1.2, model="Model.rda", file1="Coefficients.csv",
file2="Predictions.csv",  na="NA", dec=",", row.names=FALSE, file="Output.txt")
```

Arguments

`data` Data file.

`cat` Dependent variable.

`var` Independent variables.

`resp` Category that will be the reference for the dependent variable.

`stepwise` If TRUE, a logistic regression is applied by steps, in order to eliminate those variables that are not significant. The Akaike information criterion (*AIC*) is used to define what are the variables that are excluded (see section *details* of the function XI5 for more details).

`cart` If TRUE, a graph of classification and regression tree (*CART*) is created.

`minsplit` CART. It specifies the minimum number of observations that must exist in a node for the partition, that is to say, as smaller is the tree, it will have more nodes.

`minbucket` CART. It defines the minimum number of observations in a terminal node and, therefore, like the previous argument, as it is smaller, it goes more deeply into the tree, that is to say, this has more nodes and becomes more complex.

`cp` CART. It controls the complexity and saves time in the pruning process. It is (default 0.0001) fraction that should improve the adjustment indicator to continue the construction of the tree. If the improvement is less than *cp* the process stops. The user decides to delve deeper and make the tree very complex (*cp* small), or have a more simple tree (*cp* bigger).

`surrogatestyle` CART. It controls the criterion for choosing the best substitute variable for the partition on each node. If it is 0, the variable with the greatest number of items correctly classified is used, and if it is 1, the variable which has the highest percentage of elements correctly classified on the number of valid cases of that variable; the first option penalises the variables that have missing values.

`type` CART. Format of the graph: 0 shows the variables that have been selected in each partition and in the end, the predominant categories in the end-nodes, 1 is the same as the previous one but it also shows the predominant category in each partition, 2 is equal to 1 but it puts the node above and below the variable responsible for the partition, 3 the variable responsible for the partition is placed in both branches and, finally, 4 is the same as 3 but the nodes are labelled, not the branches.

`extra` CART. It allows to show additional information on nodes: 0 only shows the majority group on the node, 1 shows the number of cases of all categories, 2 shows the number of correct cases against the total, 3 shows the number of incorrect cases against the total, 4 shows the probability of all categories, 5 is the same as 4 but does not display the category that predominates in the node, 6 shows the probability of the second type that is dominant in the node, 7 is the same as 6 but does not display the category that predominates in the node, 8 shows the probability of the type that predominates in the node, and finally, 9 shows the relative probability considering observations.

`varlen` CART. It defines the number of characters of the texts: 0 uses the full text, >0 abbreviates them taking into account the number that is indicated, and <0 cuts the text of the variables taking into account, as much as possible, the indicated number to adjust the size.

`ycompress` CART. If it is TRUE, when the nodes are overlapping they move vertically to avoid this overlap.

`main` CART. Graph title.

`cex.main` CART. Font size title.

`cex` CART. Font size of the labels of the nodes.

`model` Filename with model.

`file1` CSV FILES. Filename with regression coefficients.

`file2` CSV FILES. Filename with the predictions of the logistic model.

`na` CSV FILES. Text used in the cells without data.

`dec` CSV FILES. It defines whether the comma "," or the dot "." are used as decimal separators.

`row.names` CSV FILES. Logical value which defines whether there are identifiers on the rows or a vector with a text for each of the rows.

`file` TXT FILE. Name of the output file with the results of the logistic regression.

Details

XI. REGRESSIONS

XI.2. REGRESSION MODELS FOR QUALITATIVE DEPENDENT VARIABLES

XI.2.1. Logistic regression

In the event that the dependent variable is qualitative, a suitable model is the logistic regression. It is used widely in clinical research since it allows to estimate the probability of occurrence of a process based on certain variables, allowing to evaluate the influence of the independent variables on the dependent variable, resulting in probability. The dependent variable is always qualitative but the independent variables can be continuous, discrete, categorical, dichotomous or a mixture of them all. As was the case with the regression for quantitative dependent variables there must be multicollinearity between the different independent variables and the sample observations should be independent of each other.

However, it is not required that the residuals presents a Normal distribution or the assumption of homoscedasticity (constant variance in the residuals).

The dependent variable may be dichotomous (0 if the fact does not occur and 1 if it happens) or polynomial (there are several categories), giving rise to two different types of logistic regression, the binomial or multinomial, respectively.

XI.2.1.1. Binomial logistic regression

XI.2.1.1.1. Estimation of the model

The dependent variable is:

$$Y_i = log\left(\frac{\Pi_i}{1-\Pi_i}\right)$$

where Π_i is the probability that in the case *i* the studied event occurs and $Y_i = \alpha + \beta X_i$ is the value of the dependent variable in the case *i*.

Expressed as regression:

$$\Pi_i = \frac{1}{1+e^{-(\alpha+\beta X_i)}}$$

where α and β are the coefficients of the equation.

In case of several predictor variables (independent) regression becomes:

$$\Pi_i = \frac{1}{1+e^{\left[-(\alpha+\sum_{j=1}^{k}\beta_j X_{ji})\right]}}$$

where Π_i is the probability that the studied event occurs in the case *i*, *k* is the number of predictors, α is a coefficient, β_j is the coefficient of the predictor variable *j* and X_{ji} is the value of the predictor variable *j* in the case *i*.

In the case of Π_i is greater than 0.5, it is assumed (for the purposes of prediction) that the event occurs and if this is less than 0.5, that it does not occur.

As a complement to the logistic regression, with this function the trees of classification and regression (CARTs) are used. This is a non-parametric procedure of predicting a dependent variable based on a set of independent variables (Breiman's algorithm et al., 1984). Its non-parametric character makes no assumptions regarding the distribution of the dependent and independent variables is required, nor to the relationship between them and their possible interactions.

The dependent variable can be categorical (Classification Trees), different species such as, rejection treatment, presence/absence of cancer, etc. The dependent variable may also be continuous (Regression Trees), as for example the body mass index, blood sugar, etc. In both cases, Classification and Regression Trees (CARTs),the aim is to identify the variables that best identify the dependent variable. In this function the CARTs are used to determine the variables that best discriminate the two categories of the dependent variable. They are, therefore, an excellent tool to choose the best model to help define which are the independent variables with the greatest power of discrimination. For more details see the operating manual of the function rpart and/or Guisande et al. (2012).

FUNTIONS

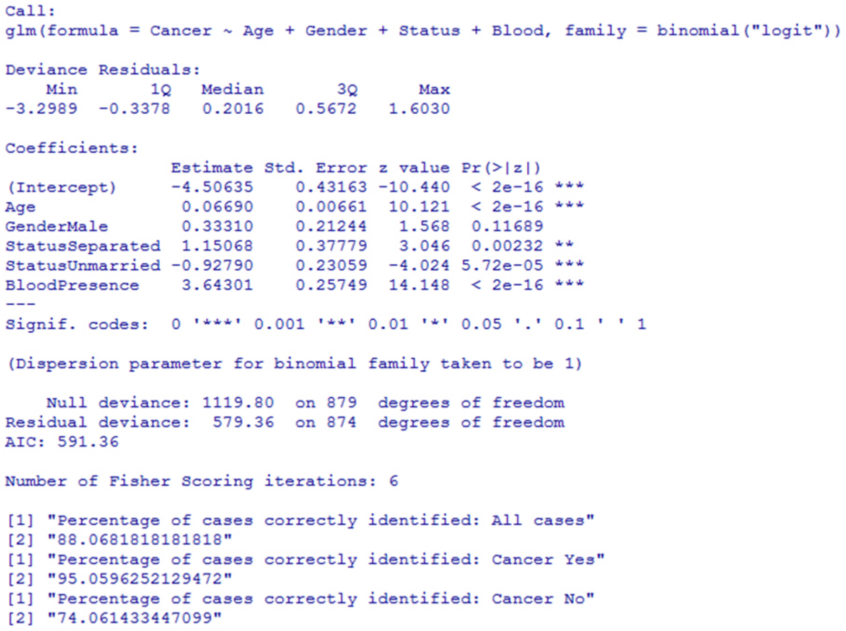
To estimate binomial logistic regression the glm function of the package stats is used. To estimate the model by steps, the stepAIC function of the package MASS (Venables & Ripley, 2002; Ripley et al., 2014) was used. To estimate the CART, the rpart function (Therneau et al., 2014) and for its representation, the rpart.plot function (Milborrow, 2014).

EXAMPLE

The study was to examine patients who have been diagnosed with a type of cancer, as well as a sample of people who do not. The objective is to determine if variables such as gender, age, marital status (unmarried, married and separated) and presence/absence of blood in the urine, can be an indicator of the presence of cancer.

The options were left by default and, therefore, the regression is performed by steps. A *minbucket= 10* was selected to do a more simplified CART, since it is only interesting to get information about what are the variables that best discriminate between the different families.

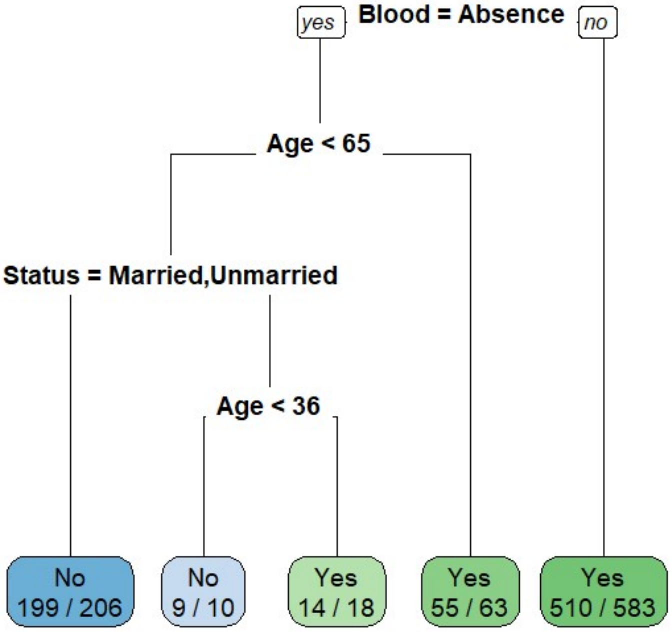
The results show the regression coefficients, which are also obtained in a CSV file. The model correctly predicts the 74.1% of the people who do not have cancer. However, what is important is that it is able to correctly predict the 95.1% of the people who have cancer. Therefore, it is a relatively good model to be able to predict when a person could potentially have such cancer.



```
Call:
glm(formula = Cancer ~ Age + Gender + Status + Blood, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2989  -0.3378   0.2016   0.5672   1.6030

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.50635    0.43163 -10.440  < 2e-16 ***
Age              0.06690    0.00661  10.121  < 2e-16 ***
GenderMale       0.33310    0.21244   1.568  0.11689
StatusSeparated  1.15068    0.37779   3.046  0.00232 **
StatusUnmarried -0.92790    0.23059  -4.024 5.72e-05 ***
BloodPresence    3.64301    0.25749  14.148  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1119.80  on 879  degrees of freedom
Residual deviance:  579.36  on 874  degrees of freedom
AIC: 591.36

Number of Fisher Scoring iterations: 6

[1] "Percentage of cases correctly identified: All cases"
[2] "88.0681818181818"
[1] "Percentage of cases correctly identified: Cancer Yes"
[2] "95.0596252129472"
[1] "Percentage of cases correctly identified: Cancer No"
[2] "74.061433447099"
```

The CART shows that the presence of blood in urine is the most important variable (Figure XI8.1). As the coefficient for the category is positive (3.64) and this coefficient refers to the category of blood in urine (presence), this indicates that people with cancer often have blood in the urine. The second variable in importance is the age. The positive coefficient (0.067) indicates that the occurrence of cancer is more frequent at higher age. Finally, the marital status of the person is incorporated in the CART (Figure XI8.1). In the event that the categorical variables are not dichotomous (only marital status), it creates a coefficient *B* for each different category of reference (which has a value of 0). The negative value of the coefficient *B* indicates that among single people (-0.927) the presence of cancer is less common. The fact that they were a man or a woman, does not influence significantly in the presence of cancer (p = 0.11), and although it is included in the model, possibly their exclusion would not mean changes in the predictive capacity of the model.

Figure XI8.1. Classification and Regression Tree that shows the importance of the variables analyzed in the study to predict the presence of cancer.



The model would be defined by the following equation:

$$Pi_{Cancer.Yes} = \frac{1}{1 + e^{-(-4.5+0.067*Age+0.333*Gender.Male+1.151*Separated-0.928*Unmarried+3.643*Blood.Presence)}}$$

For example, suppose the first case of the data table, the probability that a married man, 59, and the presence of blood in the urine, has this type of cancer would be of 0.96 and, therefore, as it is greater than 0.5, predicted that the person might have cancer:

$$Pi_{Cancer.Yes} = \frac{1}{1 + e^{-(-4.5+0.067*59+0.333*1+1.151*0-0.928*0+3.643*1)}} = 0.96$$

The probability that a single man of 23 years, and without blood in the urine, has this kind of cancer it would be 0.027 and, therefore, as it is less than 0.5, it is predicted that the individual did not have cancer:

$$Pi_{Cancer.Yes} = \frac{1}{1 + e^{-(-4.5+0.067*23+0.333*1+1.151*0-0.928*1+3.643*0)}} = 0.027$$

Value

A TXT file with the coefficients of the logistic regression and the percentage of cases correctly identified is obtained. Two CSV files are obtained: (1) the predictions of the logistic model and 2) the regression coefficients. Finally, the classification and regression tree is shown, if it has been selected.

References

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) *Classification and Regression Trees*. Wadsworth.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Milborrow, S. (2014) Plot rpart models. An enhanced version of plot.rpart. R package version 1.4-4. Available at: http://CRAN.R-project.org/package=rpart.plot.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: http://CRAN.R-project.org/package=MASS.

Therneau, T., Atkinson, B. & Ripley B (2014) Recursive Partitioning and Regression Trees. R package version 4.1-8. Available at: http://CRAN.R-project.org/package=rpart.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXI7)

XI6(data=ZXI7, cat="Cancer", var=c("Age","Gender", "Status","Blood"),
resp="Yes", minbucket=10)


## End(Not run)
```

XI7 *BINOMIAL LOGISTIC REGRESSION-PREDICTION*

Description

It allows to calculate the category to which a case belongs, being able to carry out the calculation for many cases at the same time, using a logistic regression model estimated with the binomial function XI6.

Usage

```
XI7(data, cat, resp, model="Model.rda", file="Model predictions.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Data file.
`cat`	Dependent variable.
`resp`	Category that will be the reference for the dependent variable.
`model`	The file name with the estimated model using the function XI6
`file`	CSV FILE. Filename with the predictions of the binomial logistic model, for each case.
`na`	CSV FILE. Text that is used in the cells without data.
`dec`	CSV FILE. It defines whether the comma "," or dot "." are used as decimal separators.
`row.names`	CSV FILE. Logical value which defines whether there are identifiers on the rows or a vector with a text for each of the rows.

Details

XI.2.1. Logistic regression

XI.2.1.1. Binomial Logistic Regression

XI.2.1.1.2. Prediction of the model

The percentage of cases correctly identified is an important parameter to evaluate the quality of the logistic regression model. However, a better parameter to evaluate the quality of the model is to check the degree of success using some data that had not been included within the original matrix that was used to estimate the model.

This function allows to easily predict the group to which a case belongs, which in addition to being used to evaluate the quality of the model, it also serves to be able to use the model, i.e., to predict to which category a particular case belongs, which may be of great interest for example in medicine.

In order to use this function, it is necessary, to have first calculated the logistic model with the binomial XI6 function.

EXAMPLE

The data must have the same format, that is, the same variables in the same position as the matrix of data used to estimate the model with the XI6 function. In addition, the working directory must be the same that was used when the estimated model with the XI6 function.

The study consisted of examining patients which have been diagnosed a type of cancer, as well as a sample of people who do not have it. However, these data were not included when the original binomial logistic regression model was estimated (see the example of the XI6 function). The objective is to predict whether a person can have that type of cancer.

As mentioned above, in order to use this function, it is necessary to have first calculated the binomial logistic model with the binomial logistic model with the XI6 function. It is also very important that the group of reference, in this case *resp="Yes"*, be equal to the specified one when calculating the binomial logistic model with the XI6 function.

The results show that the success rate is 88.3%, in these 762 cases that the database comprise.

Value

A CSV file is obtained with the predictions of the logistic model for each case.

Examples

```
## Not run:

data(ZXI8)

XI7(data=ZXI8, cat="Cancer", resp="Yes")


## End(Not run)
```

XI8 *MULTINOMIAL LOGISTIC REGRESSION-ESTIMATION*

Description

Different statistical techniques are applied to obtain the type of function that best fits the possible relationship between a qualitative dependent variable with multiple categories and several qualitative or quantitative independent variables.

Usage

```
XI8(data, cat, var, optimize=TRUE, methop="cart", stepwise=FALSE, order=TRUE,
cart=TRUE, minsplit=2, minbucket=2, cp=0.0001, surrogatestyle=0, type=0,
extra=2, varlen=0, ycompress= TRUE, main=NULL, cex.main=2, cex=1.2,
file1="Predictions.csv", file2="Coefficients.csv", file3="Table.csv",
na="NA", dec=",", row.names=FALSE, model= "Model.rda", file="Output.txt")
```

Arguments

`data` Data file.

`cat` Dependent variable.

`var` Independent variables.

`optimize` If TRUE, two different algorithms can be used, which are defined in the argument *methop*, whose goal is to look for a model that correctly identifies the 100% of the cases. If it is TRUE, the first variable argument of the *var* should be quantitative. If it is TRUE, the model saved, which can then be used for prediction in the function XI9, is the optimized and not the logistical nor by steps, although both are calculated.

`methop` It defines the algorithm that is used to optimize the model. If "cart", this is calculated with the variables that contribute to differentiating the categories using the trees of regression and classification (CARTs), which is explained in the function XIV1, and then those variables are introduced into the regression model in order of importance. If it is "stepwise", a stepwise logistic regression is applied to eliminate those variables that are not significant, using the information criteria of Akaike (*AIC*) to define which are the variables that are excluded (see section *details* of function XI5 for more details). After the stepwise regression, more variables are added to the model if they help increase the percentage of cases correctly identified. Therefore, it is estimated a model with all variables, the stepwise logistic regression model and the optimized model, which is the latest model that is saved.

`stepwise` If it is TRUE, a logistic regression is applied by steps, in addition to the model without steps, in order to eliminate those variables that are not significant using the AIC criterion mentioned above in the argument *methop*. The difference with the optimize method *optimize=TRUE* is that subsequently more variables are not added to the model using the approach to increase the percentage of cases correctly identified. This argument is placed automatically in TRUE if the argument *optimize=TRUE* and the *methop= "stepwise"*.

`order` If it is TRUE, the regression coefficients are ordered in the file *Coefficients.CSV* from A to Z for characters and from lesser to greater for numbers.

`cart` If TRUE, a graph of classification and regression tree (*CART*) is created.

`minsplit` CART. It specifies the minimum number of observations that must exist in a node for which the partition is made, that is to say, as smaller the tree will have more nodes.

`minbucket` CART. It defines the minimum number of observations in a terminal node and, therefore, like the previous argument, as it is smaller, it goes more deeply into the tree, i.e. has more nodes and is more complex.

`cp` CART. Controls the complexity and allows you to save time in the process of pruning. It it is the fraction (by default 0.0001) in which this should improve the adjustment indicator to continue the construction of the tree. If the improvement is less than *cp* the process stops. The user decides whether to delve deeper and make the tree very complex (*cp* small), or have a more simple tree (*cp* bigger).

`surrogatestyle` CART. It controls the criterion for choosing the best surrogate variable for the partition on each node. If it is 0, the variable with the largest number of items correctly classified is used, and if it is 1, the one that has the highest percentage of elements correctly classified on the number of valid cases of that variable; the first option penalises the variables that have missing values.

`type` CART. Format of the chart: 0 shows the variables that have been selected in each partition and in the end, the predominant categories in the end-nodes, 1 is the same as the previous one but it also shows the predominant category in each partition, 2 equal to 1 but it puts the node above and below the variable responsible for the partition, 3 the variable responsible for the partition is placed in both branches and, 4 is the same as 3 but the nodes are labelled, not the branches.

`extra` CART. It allows to display additional information in the nodes: 0 shows only the majority group in the node, 1 shows the number of cases of all categories, 2 shows the number of correct cases compared to the total, 3 shows the number of incorrect cases compared with the total, 4 shows the probability of all the categories, 5 is equal to 4 but it does not display the category that predominates in the node, 6 shows the probability of the second class that is dominant in the node, 7 is the same as 6 but it does not display the category which is dominant in the node, 8 shows the probability of the class that predominates in the node, and finally, 9 shows the relative likelihood considering all the observations.

`varlen` CART. It defines the number of characters of the texts: 0 uses the full text, >0 uses the abbreviated one taking into account the number that is indicated, and <0 clips the text from the variables taking into account, as much as possible, the indicated number to adjust the size.

`ycompress` CART. If TRUE, when the nodes are overlapping these move vertically to avoid this overlap.

`main` CART. Graph title.

`cex.main` CART. Font size title.

`cex` CART. Font size of the labels of the nodes.

`file1` CSV FILE. Filename with predictions for each case of the logistic model, stepwise model and the optimized model, if that was selected to perform the last two.

`file2` CSV FILES. Filename with regression coefficients. If *optimize=TRUE* are the coefficients of the model optimized, if *optimize=FALSE* and *stepwise=TRUE* are the coefficients of the model by steps and, finally, if *optimize=FALSE* and *stepwise=FALSE* are the coefficients of the model in which all the variables are included.

`file3` CSV FILES. Filename with the percentage of cases identified for each category of the dependent variable.

`na` CSV FILES. Text that is used in the cells without data.

`dec` CSV FILES. It defines whether the comma "," or dot "." are used as decimal separators.

`row.names` CSV FILES. Logical value which defines whether there are identifiers on the rows or a vector with a text for each of the rows.

`model` Filename with model.

`file` TXT FILE. Output file name with the results.

Details

XI.2.1. Logistic regression

XI.2.1.2. Multinomial Logistic Regression

XI.2.1.2.1. Estimation Model

The main difference with respect to the binomial logistic regression is that qualitative dependent variable is not dichotomous, but may have more than 2 categories (for details see Guisande et al., 2011). Assuming *k* classes or categories, and *j* independent variables, this model can be summarized in the following functions, which provide the probability of belonging to the first *k* - 1 classes:

$$Z_{in} = \beta_{n0} + \beta_{n1}X_{i1} + \beta_{n2}X_{i2} + ... + \beta_{nj}X_{ij}$$

$$\Pi_{in} = \frac{e^{Z_{in}}}{1 + e^{Z_{i1}} + e^{Z_{i2}} + e^{Z_{i3}} + ... + e^{Z_{ik-1}}}$$

where Π_{in} is the belonging probability of the case *i* to the class *n*; Z_{in} is the value of the dependent variable Z corresponding to the n class in the case *i*; β_{nj} is the coefficient of the independent variable *j* for the *n* class; x_{ij} is the value of the predictor or independent variable *j* for the case *i*. The probability for the last class *k* is obtained from the difference to 1.

As in the case of binomial logistic regression, classification and regression tree (CART) is an excellent tool to identify which are the variables that have the greatest differences between the categories of the dependent variable (based on Breiman's Algorithm et al., 1984) and, therefore, this makes it easier to choose the variables that going to get a greater number of cases correctly identified with the regression model. Therefore, this is helpful for obtaining the most predictive model of logistic regression. For more details, see the manual of the rpart function and/or Guisande et al. (2012).

FUNCTIONS

For the estimation of multinomial logistic regression the multinom function is used (Ripley, 2014). To estimate the model by steps the stepAIC function of the MASS package was used (Venables & Ripley 2002; Ripley et al. , 2014). For the estimation of the CART, the rpart function was used (Therneau et al. , 2014) and for their representation the rpart.plot function was used (Milborrow, 2014).

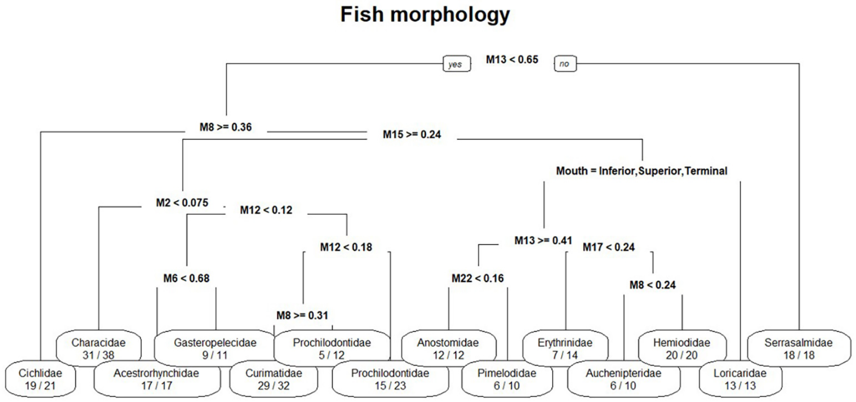
EXAMPLE

The data are body measures of several species of fish (continuous variables), as well as the type of mouth they have (categorical variable). The objective is to determine, if it is possible to identify, which family an individual belongs in function of biometrics.

Step 1.

First the dependent variable *cat="Family"* and all morphometric as independent variables are selected. Default conditions are not neglected, since the argument *optimize=FALSE* and, therefore, not optimized model is made. Conversely the argument *stepwise=TRUE* and, therefore, the regression is performed by steps. A *minbucket=10* was chosen for a more simplified CART, since only matter which are the variables that best discriminate between different families.

Figure XI8.2. Classification and Regression Tree of measures of freshwater fishes.



The CART obtained (Figure XI8.2) shows the most important variables that best discriminate against families and this is even a better method for selecting variables than stepwise regression. The variables that are higher up the tree and are responsible for the first nodes are those which have a greater capacity for discrimination.

In the logistic model without steps, which includes all the variables, a 100% of cases correctly identified is achieved, while in the model by steps a 97.6% success is achieved. In the CSV file that is generated to save the estimated coefficients, whose default name is *Coefficients.CSV*, it is noted that the selected variables in the model by steps are M12, M16, M17 and M22. There are no coefficients for the first category, which is calculated by default from the other categories. In the *Predictions.CSV* file, the actual and predicted categories for each case are shown by the models without and by steps.

Step 2.

Then, a model in which it is the default otions is made, i.e., the *optimize=TRUE* argument. On the other hand, the CART *cart=FALSE* is not performed. The results show that the model correctly identifies the 100% of the cases.

Once developed the model it is possible to use it to predict which category an individual belongs. The XI9 function allows to calculate the category to which an individual belongs, being able to carry out the calculation for many individuals at once, using the regression model without steps, by steps or optimized, which are estimated with this function.

Value

A TXT file is obtained with the coefficients of the logistic regression that it is not by steps, the percentage of cases correctly identified, and the coefficients of the logistic regression by steps and the percentage of cases correctly identified, in the case that the regression by steps is selected. Five CSV files are obtained: (1) the predictions of the logistic model for each case and the model by steps, in the case of selecting it; (2) the regression coefficients; 3) the regression coefficients by steps, if this is chosen; 4) a table with the percentage of cases identified for each group and 5) a table with the percentage of cases identified for each group using the logistic model by steps, if this was selected. It also generates a file with the logistic model without steps and another for the model by steps, if this was selected, which are used later in the XI9 function to predict to what categories the individuals belong. Finally, it shows the classification and regression tree, if this has been selected.

References

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) *Classification and Regression Trees*. Wadsworth.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Milborrow, S. (2014) Plot rpart models. An enhanced version of plot.rpart. R package version 1.4-4. Available at: `http://CRAN.R-project.org/package=rpart.plot`.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: `http://CRAN.R-project.org/package=MASS`.

Ripley, B. (2014) Software for feed-forward neural networks with a single hidden layer, and for multinomial log-linear models. R package version 7.3-8. Available at: `http://CRAN.R-project.org/package=nnet`.

Therneau, T., Atkinson, B. & Ripley B (2014) Recursive Partitioning and Regression Trees. R package version 4.1-8. Available at: http://CRAN.R-project.org/package=rpart.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXI9)

#Step 1

windows(26,14)

XI8(data = ZXI9, cat="Family", var=c("M2","M6","M8","M9", "M12","M13",
"M15","M16","M17","M21","M22","Mouth"), optimize=FALSE, stepwise=TRUE,
minbucket=10, main="Fish morphology", cex=1)

#Step 2

XI8(data=ZXI9, cat="Family", var=c("M2","M6","M8","M9", "M12","M13",
"M15","M16","M17","M21","M22","Mouth"), cart=FALSE)


## End(Not run)
```

XI9 *MULTINOMIAL LOGISTIC REGRESSION-PREDICTION*

Description

It allows to calculate the category to which a case belongs, being able to carry out the calculation for many cases at the same time, using the regression models without steps, by steps or optimized, which are estimated with the XI8 function.

Usage

```
XI9(data, cat, model="Model.rda", file="Model predictions.csv", na="NA",
dec=",", row.names=FALSE)
```

Arguments

data	Data file.
cat	Dependent variable.
model	The file name with the estimated model using the function XI8.
file	CSV FILE. Filename with the predictions of the multinomial logistic model, for each case.

`na` CSV FILE. Text that is used in the cells without data.

`dec` CSV FILE. It defines whether the comma "," or dot "." are used as decimal separators.

`row.names` CSV FILE. Logical value which defines whether there are identifiers on the rows or a vector with a text for each of the rows.

Details

XI.2.1. Logistic regression

XI.2.1.2. Multinomial Logistic Regression

XI.2.1.2.2. Prediction of the model

The percentage of cases correctly identified is an important parameter to evaluate the quality of the logistic regression model. However, a better parameter to evaluate the quality of the model is to check the degree of success using some data that had not been included within the original matrix used to estimate the model.

This function allows to easily predict the group to which a case belongs, which in addition to being used to evaluate the quality of the model, it also serves to be able to use the model, i.e., to predict to which category a particular case belongs, which may be of great interest for example in medicine.

In order to use this function, it is necessary, to have first calculated the logistic model with the multinomial XI8 function.

EXAMPLE

The data must have the same format, that is, the same variables in the same position as the matrix of data used to estimate the model with the XI8 function. In addition, the working directory must be the same that was used when the estimated model with the XI8 function.

The data are body measures of several species of fish (continuous variables), as well as the type of mouth they have (categorical variable), of the same population of the data that were used for the example of the function XI8. However, these data were not included when the original model was estimated. As it is known to which family each individual belongs, the objective is to evaluate the quality of the model by means of the degree of identification success with these individuals.

As mentioned above, it is necessary to have first calculated the regression models without steps, by steps or optimized, which are estimated with the XI8 function.

The results show that the success rate is 100%, in these 29 individuals that the database comprise. Therefore, it seems that the model obtained is quite reliable.

Value

A CSV file is obtained with the predictions of the logistic model for each case.

Examples

```
## Not run:

data(ZXI10)

XI9(data=ZXI10, cat="Family")
```

```
## End(Not run)
```

XII1 *META-ANALYSIS/FOREST PLOT*

Description

A meta-analysis is applied and shown in a forest plot.

Usage

```
XII1(data, event.e, n.e, event.c, n.c, studlab, META=NULL,
FOREST=NULL, file=”Output.txt”)
```

Arguments

`data`	Data file.
`event.e`	Variable with the number of events in the experimental group.
`n.e`	Variable with the total number of observations in the experimental group.
`event.c`	Variable with the number of events in the control group.
`n.c`	Variable with the total number of observations in the control group.
`studlab`	Variable with the code of the scientific studies.
`META`	It accesses the metabin function which allows to modify the arguments of the meta-analysis.
`FOREST`	It accesses the function forest.meta which allows to modify the arguments of the graph of the meta-analysis.
`file`	TXT FILE. Name of the output file with the results.

Details

XII. META-ANALYSIS

XII.1. FOREST PLOT

The meta-analysis is a statistical technique that allows to combine several statistical studies in only one. With this, statistical summaries are constructed so that they can take advantage of the increased power provided by the sample size, necessarily larger, resulting from the aggregation of the different samples of all the studies. This analysis is of special interest in medicine, where it is important to determine if a particular medication, a compound, a therapy, etc., can be beneficial or detrimental to health, especially in those cases in which there is a discrepancy between the different studies. As previously mentioned, the advantage of the meta-analysis is that it allows to combine the information from all the work done. The representations that are most frequently used are the forest plot, and the diagram of funnel.

FUNCTIONS

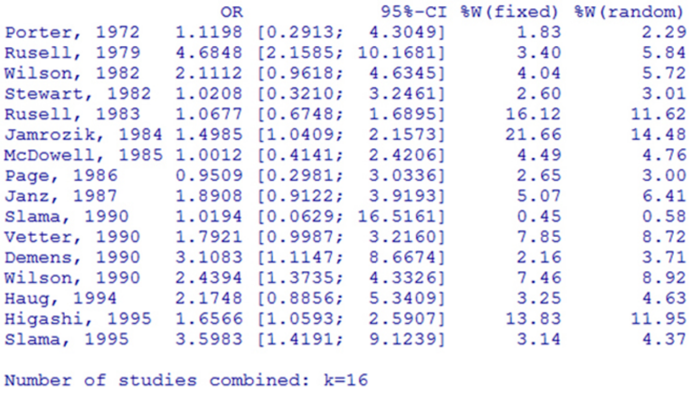
The meta-analysis will use the metabin function and to the forest plot forest.meta, both from the meta package (Schwarzer, 2013). For more details on how to use the arguments to these functions, which gives access to the arguments *META* and *FOREST*, consult the help of the function and/or Guisande & Vaamonde (2012).

EXAMPLE

These data correspond to a study of Silagy & Ketteridge (1997) on the effect that the medical recommendation has to quit smoking. In the data file, the variable *event.e* refers to the number of people who stopped smoking based on the recommendation of the doctor, *n.e* the total number of people in the group to which the doctor recommended stop smoking, and *event.c* and *n.c* are the people who had stopped smoking and the total number of people who were in the control group, to which the doctor, did not recommend quitting tobacco, respectively.

In the function metabin, the argument *method* specifies the method of grouping cases: "MH" Mantel-Haenszel , "Inverse" inverse variance, or "Peto" Peto method. Mantel-Haenszel (default option) calculates the aggregate OR weighting the data from each study by the total number of items that it contains considering its distribution into classes, and is the most widely used. Peto is an alternative method for the calculation of the odds that does not require correction for cells with zero cases, so it can be used when investigating rare events and the sample size is small.

The first results observed in the output TXT file are shown below:



```
                   OR             95%-CI %W(fixed) %W(random)
Porter, 1972   1.1198 [0.2913;  4.3049]       1.83       2.29
Rusell, 1979   4.6848 [2.1585; 10.1681]       3.40       5.84
Wilson, 1982   2.1112 [0.9618;  4.6345]       4.04       5.72
Stewart, 1982  1.0208 [0.3210;  3.2461]       2.60       3.01
Rusell, 1983   1.0677 [0.6748;  1.6895]      16.12      11.62
Jamrozik, 1984 1.4985 [1.0409;  2.1573]      21.66      14.48
McDowell, 1985 1.0012 [0.4141;  2.4206]       4.49       4.76
Page, 1986     0.9509 [0.2981;  3.0336]       2.65       3.00
Janz, 1987     1.8908 [0.9122;  3.9193]       5.07       6.41
Slama, 1990    1.0194 [0.0629; 16.5161]       0.45       0.58
Vetter, 1990   1.7921 [0.9987;  3.2160]       7.85       8.72
Demens, 1990   3.1083 [1.1147;  8.6674]       2.16       3.71
Wilson, 1990   2.4394 [1.3735;  4.3326]       7.46       8.92
Haug, 1994     2.1748 [0.8856;  5.3409]       3.25       4.63
Higashi, 1995  1.6566 [1.0593;  2.5907]      13.83      11.95
Slama, 1995    3.5983 [1.4191;  9.1239]       3.14       4.37

Number of studies combined: k=16
```

The value of *OR* (*Odds Ratio*), which measures the effect of the doctor's advice to quit smoking is calculated as follows, for example, in the case of the second study of Russell in 1979:

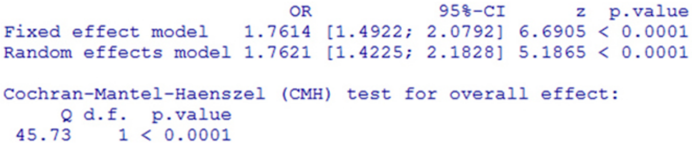
$$OR = \frac{34/(1031-34)}{8/(1107-8)} = 4.68$$

This value indicates that among those patients who received the medical advice, the relationship between those who stopped smoking and those who did not (predominance or advantage) is 4.68 times higher than among those who did not receive medical advice. The medical advice allowed, therefore, multiply by something more than four the success reached in smoking cessation. If the OR is significantly less than one, this indicates that the studied event has a lower predominance in the experimental group than in the control group: for example, if the studied event is the death of the patient, an odds ratio less than one corresponds to a factor of protection, and one or greater than the unit to a risk factor. It is noted that not all studies have an OR as high as that of the second study calculated above; some show values next to the unit (for example the McDowell 1985 and Slama

1990), indicating a null effect of medical advice. The confidence interval of 95% that is shown in the two columns below helps us to understand better the OR: if the interval contains the value 1 the effect is not significant (i.e., it indicates that the medical advice does not appear to be effective), it occurs in 10 of the 16 studies.

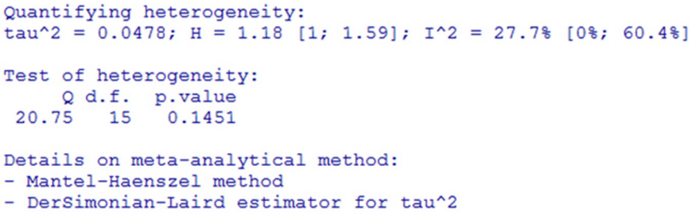
The following results that appear in the TXT file, shown below, are the weight or weighting of each study, depending on the sample sizes used in each one. The fixed effects model assumes that the studies analyzed are all that are going to be considered, and the random effects that the data used are a random sample of studies obtained from a larger set, which adds an additional source of variability. The aggregated results show: OR = 1.76; although some studies do not seem to show a real effect of the medical recommendation, this set as a whole indicates a significant effect (the interval does not contain the value 1 and the p-value of the contrast of invalidity (null hypothesis: OR = 1) is very small, less than 0.0001.

In the test of Cochran-Mantel-Haenszel (see Guisande et al. 2011, for more details on this test), the value of p < 0.001 and, therefore the number of patients who stop smoking, is different in the experimental group (advised by the doctor) in comparison with the control (not advised by the doctor). Medical advice multiplies the predominance of quitting smoking by 1.76 (ratio between those quitting and those who do not).



```
                            OR            95%-CI      z  p.value
Fixed effect model   1.7614 [1.4922; 2.0792] 6.6905 < 0.0001
Random effects model 1.7621 [1.4225; 2.1828] 5.1865 < 0.0001

Cochran-Mantel-Haenszel (CMH) test for overall effect:
      Q d.f.  p.value
 45.73    1 < 0.0001
```

The latest results show some measures of heterogeneity. When studies are very heterogeneous, that is to say very different from each other, it is questionable whether they can be aggregated into one. For this reason different measures of heterogeneity are used. These include Tau square, variability among studies, or the statistical I-squared, which measures the proportion of the total variability that is attributable to heterogeneity between studies, i.e., the variability among studies divided by the total variability = variability among studies + variability between patients. The authors of the statistical I-squared (Higgins et al., 2003) proposed the cut-off values 25 and 75 to interpret that heterogeneity is low (less than 25%), moderate (25% - 75%) or high (more than 75%). In the example, 27.6% is moderate-low and, therefore, it is not a problem. Finally, it shows the P-value of the contrast of invalidity. A value, such as p = 0.1451 observed in the example, greater than the usual level of significance 0.05 allows to accept that there is no significant heterogeneity.



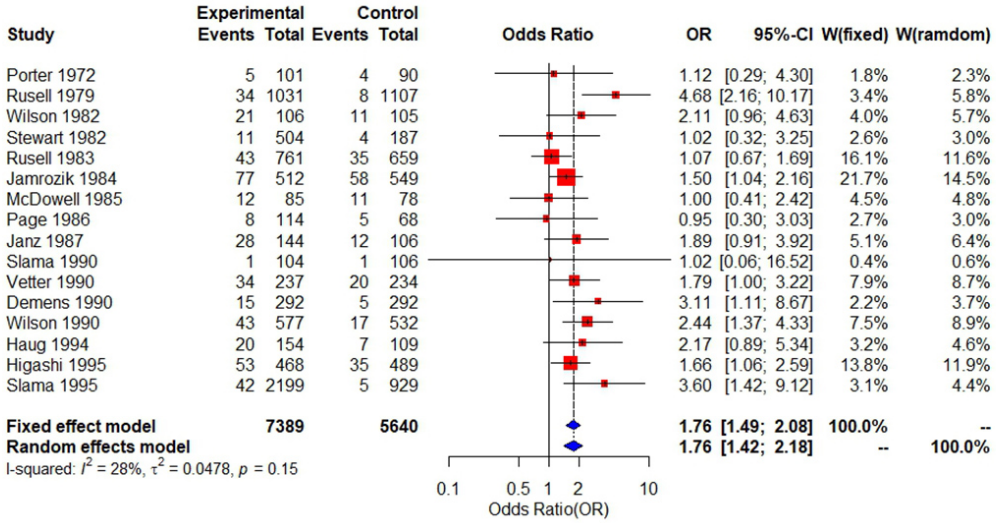
```
Quantifying heterogeneity:
tau^2 = 0.0478; H = 1.18 [1; 1.59]; I^2 = 27.7% [0%; 60.4%]

Test of heterogeneity:
      Q d.f.  p.value
 20.75   15   0.1451

Details on meta-analytical method:
- Mantel-Haenszel method
- DerSimonian-Laird estimator for tau^2
```

The forest plot (Figure XII.1) shows all the results mentioned above. The estimated value is represented by a square, and the confidence limits are the ends of each horizontal line. It can also be observed as many lines exceed the value 1 to the right, indicating that the proportion of people who quit smoking is greater in the experimental group who received medical advice, than in the control group who was not influenced by the doctor's advice. The size of the square indicates the sample used: the bigger the square more representative is the study and, therefore, more reliable their conclusions.

W is the weight or weighting that corresponds to each study on the obtaining of the *OR* average or summary. Its value depends on the aggregation method used, the sample sizes and the number of events in each study. If the studies used in the meta-analysis are all existing studies, the fixed effects model should be used and, if this is a sample of studies chosen at random, the random model should be used.

The *OR* summary calculated with the meta-analysis is represented by a diamond at the bottom, the ends of the diamond the confidence limits. It is observed as the confidence interval is much smaller than any of the individual studies, which is due to the larger size of the aggregate sample, and the fact that it is greater than 1, in particular 1.76 for both the fixed-effects and for the random, meaning that the doctor's advice positively influences at the time of quitting tobacco and, in addition, as shown in the test of Cochran-Mantel-Haenszel, this difference is significant. The graph allows, therefore, to see the results and conclusions of each of the individual studies and the aggregate results achieved with the meta-analysis.

Figure XII.1. Forest plot from a study conducted on the effect of the doctor's advice to quit smoking.



	Experimental		Control					
Study	Events	Total	Events	Total	OR	95%-CI	W(fixed)	W(ramdom)
Porter 1972	5	101	4	90	1.12	[0.29; 4.30]	1.8%	2.3%
Rusell 1979	34	1031	8	1107	4.68	[2.16; 10.17]	3.4%	5.8%
Wilson 1982	21	106	11	105	2.11	[0.96; 4.63]	4.0%	5.7%
Stewart 1982	11	504	4	187	1.02	[0.32; 3.25]	2.6%	3.0%
Rusell 1983	43	761	35	659	1.07	[0.67; 1.69]	16.1%	11.6%
Jamrozik 1984	77	512	58	549	1.50	[1.04; 2.16]	21.7%	14.5%
McDowell 1985	12	85	11	78	1.00	[0.41; 2.42]	4.5%	4.8%
Page 1986	8	114	5	68	0.95	[0.30; 3.03]	2.7%	3.0%
Janz 1987	28	144	12	106	1.89	[0.91; 3.92]	5.1%	6.4%
Slama 1990	1	104	1	106	1.02	[0.06; 16.52]	0.4%	0.6%
Vetter 1990	34	237	20	234	1.79	[1.00; 3.22]	7.9%	8.7%
Demens 1990	15	292	5	292	3.11	[1.11; 8.67]	2.2%	3.7%
Wilson 1990	43	577	17	532	2.44	[1.37; 4.33]	7.5%	8.9%
Haug 1994	20	154	7	109	2.17	[0.89; 5.34]	3.2%	4.6%
Higashi 1995	53	468	35	489	1.66	[1.06; 2.59]	13.8%	11.9%
Slama 1995	42	2199	5	929	3.60	[1.42; 9.12]	3.1%	4.4%
Fixed effect model		**7389**		**5640**	**1.76**	**[1.49; 2.08]**	**100.0%**	**--**
Random effects model					**1.76**	**[1.42; 2.18]**	**--**	**100.0%**

I-squared: $I^2 = 28\%$, $\tau^2 = 0.0478$, $p = 0.15$

Value

A TXT file is obtained with the results of the meta-analysis, in addition to the diagram of forest.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) Tratamiento de datos con R, STATISTICA y SPSS. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) Gráficos estadísticos y mapas con R. Ediciones Díaz de Santos, Madrid, 367 pp.

Higgins JPT, Thompson SG, Deeks JJ y Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ*: 327-557.

Schwarzer, G. (2013) Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot. R package version 3.5-0. Available at: `http://CRAN.R-project.org/package=meta`.

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editores. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

Examples

```
## Not run:

data(ZXII1)

#Influence of the medical advice to quit smoking.

XII1(data=ZXII1, event.e="event.e", n.e="n.e", event.c="event.c", n.c="n.c", studlab="Research",
FOREST=c("text.fixed ='Fixed effect model'",
"text.random='Random effect model'", "xlab='Odds ratio (OR)'",
"leftlabs=c('Study', 'Events', 'Total', 'Events', 'Total')",
"rightlabs=c('OR', '95
"hetlab='l-squared: '", "col.square='red'","col.diamond='blue'"))

## End(Not run)
```

XII2 *META-ANALYSIS/CUMULATIVE FOREST PLOT*

Description

A meta-analysis is applied and shown in a cumulative forest plot.

Usage

```
XII2(data, event.e, n.e, event.c, n.c, studlab, META=NULL,
FOREST=NULL, file="Output.txt")
```

Arguments

`data` Data file.

`event.e` Variable with the number of events in the experimental group.

`n.e` Variable with the total number of observations in the experimental group.

`event.c` Variable with the number of events in the control group.

`n.c` Variable with the total number of observations in the control group.

`studlab` Variable with the code of the scientific studies.

`META` It accesses the metabin function which allows to modify the arguments of the meta-analysis.

`FOREST` It accesses the function forest.meta which allows to modify the arguments of the graph of the meta-analysis.

`file` TXT FILE. Name of the output file with the results.

Details

XII. META-ANALYSIS

XII.2. CUMULATIVE FOREST PLOT

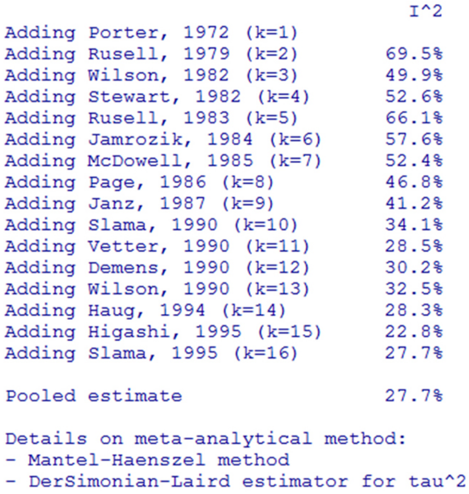
Another way to represent the forest plot (see section of *details* of the function XII1) is through the cumulative forest plot. The difference with the forest plot is that studies are being added as these began to appear, which allows to see if there is strength in the information and determine if it was a balance in the result.

FUNCTIONS

For the meta-analysis the function metabin was used, for the cumulative forest plot metacum and forest.meta, were used all of the meta package (Schwarzer, 2013). For more details on how to use these functions, which provides access from the arguments in META and FOREST, refer to the help section of the function and/or Guisande & Vaamonde (2012).

EXAMPLE

The example is the same as shown in the section of *details* of the XII1 function.



```
                                     I^2
Adding Porter, 1972 (k=1)
Adding Rusell, 1979 (k=2)          69.5%
Adding Wilson, 1982 (k=3)          49.9%
Adding Stewart, 1982 (k=4)         52.6%
Adding Rusell, 1983 (k=5)          66.1%
Adding Jamrozik, 1984 (k=6)        57.6%
Adding McDowell, 1985 (k=7)        52.4%
Adding Page, 1986 (k=8)            46.8%
Adding Janz, 1987 (k=9)            41.2%
Adding Slama, 1990 (k=10)          34.1%
Adding Vetter, 1990 (k=11)         28.5%
Adding Demens, 1990 (k=12)         30.2%
Adding Wilson, 1990 (k=13)         32.5%
Adding Haug, 1994 (k=14)           28.3%
Adding Higashi, 1995 (k=15)        22.8%
Adding Slama, 1995 (k=16)          27.7%

Pooled estimate                    27.7%

Details on meta-analytical method:
- Mantel-Haenszel method
- DerSimonian-Laird estimator for tau^2
```

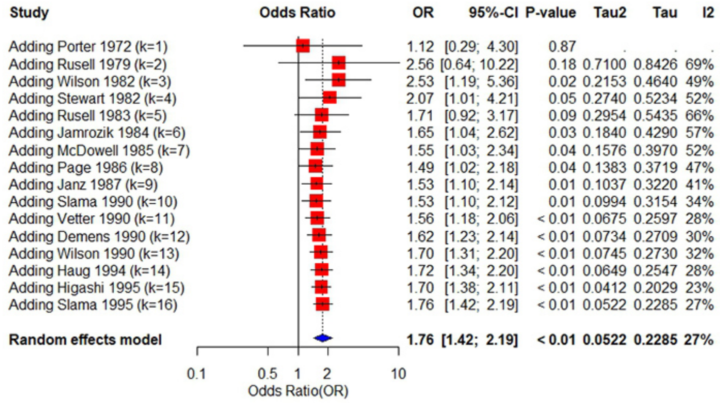
The first results showing the TXT file have already been explained in the section *details* of the function XII1. The following table shows the values of I^2, which has already been explained in the section *details* of the function XII1. The authors of the statistical I-squared (Higgins et al., 2003) proposed a heterogeneity moderate if the values are between the 25% -75%. In our example this is stabilized at the end with a value of 27.7%, which is moderate-low and, therefore, is not a problem.

The odds and confidence intervals for the different studies are displayed in the cumulative forest plot that is obtained (Figure XII.2). The estimated value is represented by a square, and the confidence limits are the ends of each horizontal line. It is observed as many lines exceed the value 1 to the right, indicating that the proportion of people who quit smoking is greater in the experimental group who was advised by the physician, than in the control group who was not influenced by the recommendations of the doctor. The size of the square indicates the sample used: the bigger the square more representative is the study and, therefore, more reliable their conclusions.

In this representation the studies are added as they were appearing, which allows to see if there is strength in the information and, if there was a balance in the result, when this balance was reached. Although the chronological order is the method most often used when studies are added to the meta-analysis, it is possible to use any other order.

The graph shows how aggregate estimation is stabilized, after an initial period with remarkable fluctuation, so that from 1990, the successive studies do not provide significant changes, and the result of the meta-analysis is virtually the same.

Figure XII.2. Cumulative forest plot from a study conducted on the effect of the doctor's advice to quit smoking.



Study	OR	95%-CI	P-value	Tau2	Tau	I2
Adding Porter 1972 (k=1)	1.12	[0.29; 4.30]	0.87	.	.	.
Adding Rusell 1979 (k=2)	2.56	[0.64; 10.22]	0.18	0.7100	0.8426	69%
Adding Wilson 1982 (k=3)	2.53	[1.19; 5.36]	0.02	0.2153	0.4640	49%
Adding Stewart 1982 (k=4)	2.07	[1.01; 4.21]	0.05	0.2740	0.5234	52%
Adding Rusell 1983 (k=5)	1.71	[0.92; 3.17]	0.09	0.2954	0.5435	66%
Adding Jamrozik 1984 (k=6)	1.65	[1.04; 2.62]	0.03	0.1840	0.4290	57%
Adding McDowell 1985 (k=7)	1.55	[1.03; 2.34]	0.04	0.1576	0.3970	52%
Adding Page 1986 (k=8)	1.49	[1.02; 2.18]	0.04	0.1383	0.3719	47%
Adding Janz 1987 (k=9)	1.53	[1.10; 2.14]	0.01	0.1037	0.3220	41%
Adding Slama 1990 (k=10)	1.53	[1.10; 2.12]	0.01	0.0994	0.3154	34%
Adding Vetter 1990 (k=11)	1.56	[1.18; 2.06]	< 0.01	0.0675	0.2597	28%
Adding Demens 1990 (k=12)	1.62	[1.23; 2.14]	< 0.01	0.0734	0.2709	30%
Adding Wilson 1990 (k=13)	1.70	[1.31; 2.20]	< 0.01	0.0745	0.2730	32%
Adding Haug 1994 (k=14)	1.72	[1.34; 2.20]	< 0.01	0.0649	0.2547	28%
Adding Higashi 1995 (k=15)	1.70	[1.38; 2.11]	< 0.01	0.0412	0.2029	23%
Adding Slama 1995 (k=16)	1.76	[1.42; 2.19]	< 0.01	0.0522	0.2285	27%
Random effects model	**1.76**	**[1.42; 2.19]**	**< 0.01**	**0.0522**	**0.2285**	**27%**

Value

a TXT file is obtained with the results of the meta-analysis, in addition to the diagram of forest.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Higgins JPT, Thompson SG, Deeks JJ y Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ*: 327-557.

Schwarzer, G. (2013) Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=meta.

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editors. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

Examples

```
## Not run:

data(ZXII1)

#Influence of the medical advice to quit smoking.

XII2(data=ZXII1, event.e=”event.e”, n.e=”n.e”, event.c=”event.c”, n.c=”n.c”,
studlab=”Research”,
FOREST=c(”text.random='Random effect model'”,
”xlab='Odds ratio (OR)'”, ”leftlabs=c('Study')”,
”col.square='red'”,”col.diamond='blue'”))



## End(Not run)
```

XII3 *META-ANALYSIS/L'ABBE PLOT*

Description

A meta-analysis is applied and displayed in a graph of L'Abbé.

Usage

```
XII3(data, event.e, n.e, event.c, n.c, studlab, META=NULL, LABBE=NULL,
file=”Output.txt”)
```

Arguments

`data`	Data file.
`event.e`	Variable with the number of events in the experimental group.
`n.e`	Variable with the total number of observations in the experimental group.
`event.c`	Variable with the number of events in the control group.
`n.c`	Variable with the total number of observations in the control group.
`studlab`	Variable with the code of the scientific studies.
`META`	It accesses the metabin function which allows to modify the arguments of the meta-analysis.
`LABBE`	It accesses the function labbe.metabin which allows to modify the arguments of the L'Abbé graph.
`file`	TXT FILE. Name of the output file with the results.

Details

XII. META-ANALYSIS

XII.3. L'ABBÉ GRAPH

In the event that the heterogeneity is present in the studies of a meta-analysis (see section of *details* of the XII1 function), the graph of L'Abbé allows to identify which are the studies responsible for this heterogeneity.

The graph of L'Abbé helps deepen the analysis of the presence and causes of heterogeneity in meta-analysis. The outcome in the treatment group (*axis y*) compared to results in the control group (*axis x*), with a diagonal line to 45° that divides the graphic into two parts: on one side are studies in which the experimental group was favorable, and in the other studies in which the control group was favorable.

FUNCTIONS

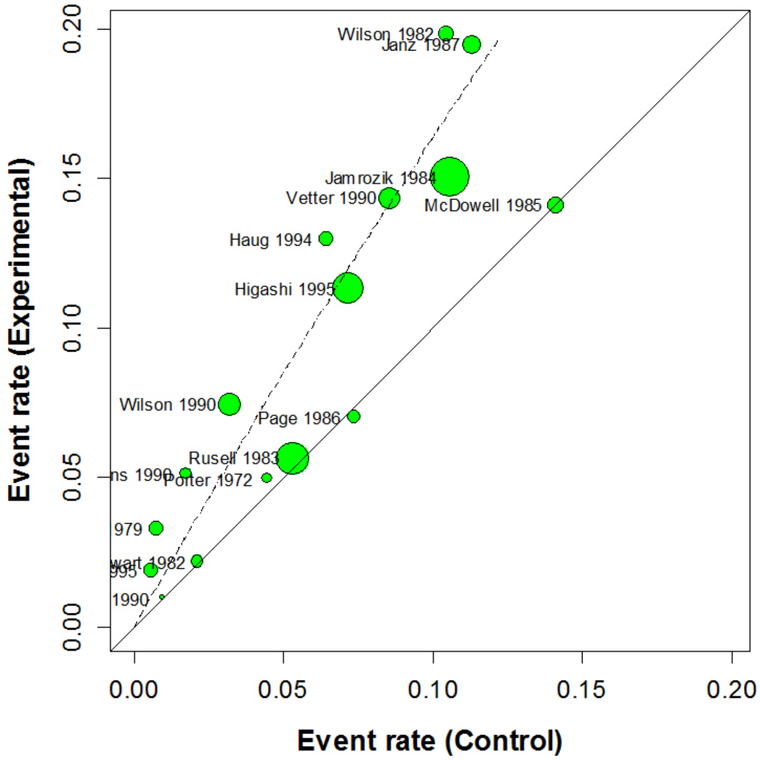
For the meta-analysis the function metabin is used for the graph of L'Abbé labbe.metabin, both from meta package (Schwarzer, 2013). For more details on how to use these functions, which provides access from the arguments META and LABBE, refer to the help of the function and/or Guisande & Vaamonde (2012).

EXAMPLE

The example is the same as it was explained in detail in the section of *details* of the function XII1.

All the results of the TXT file were also explained in the section of *details* of the function XII1.

Figure XII.3. L'Abbé plot from a study conducted on the effect of the doctor's advice to quit smoking.



In the graph of L'Abbé of the example (Figure XII.3) above the diagonal are studies in which the proportion of patients who stopped smoking thanks to medical advice is greater than in the control group. The points are plotted in varying sizes depending on the sample size and precision of each estimate, so that larger dots are the most representative studies.

If the points are grouped in a narrow zone, next to a straight line, this means that the results are homogeneous, and if they are dispersed, they show heterogeneity. It also represents the line of fit to the point cloud (dotted line), whose slope is related to a joint *OR*. It is easy to identify studies that are farthest from the majority pattern, i.e., the line of fit, which would be responsible for the observed heterogeneity, when it exists.

Value

A TXT file is obtained with the results of the meta-analysis, in addition to the L'Abbé graph.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Higgins JPT, Thompson SG, Deeks JJ y Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ*: 327-557.

Schwarzer, G. (2013) Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot. R package version 1.1-12. Available at: http://CRAN.R-project.org/package=meta.

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editors. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

Examples

```
## Not run:

data(ZXII1)

#Influence of the medical advice to quit smoking.

XII3(data=ZXII1, event.e="event.e", n.e="n.e", event.c="event.c", n.c="n.c",
studlab="Research", LABBE=c("xlim=c(-0.025,0.2)", "xlab='Event rate (Control)'",
"ylab='Event rate (Experimental)'", "main='Effect of the advice of the\
doctor to quit smoking'", "cex.main=1.4","bg='green'", "cex.lab=1.3",
"font.lab=2", "studlab=TRUE"))


## End(Not run)
```

`XII4` *META-ANALYSIS/FUNNEL PLOT*

Description

A meta-analysis is applied and shown in a funnel plot.

Usage

```
XII4(data, event.e, n.e, event.c, n.c, studlab, META=NULL, FUNNEL=NULL,
LEGEND=NULL, COLOR=c("#FFF68FFF", "#FFA54FFF", "#FF6347FF"),
file="Output.txt")
```

Arguments

`data` Data file.

`event.e` Variable with the number of events in the experimental group.

`n.e` Variable with the total number of observations in the experimental group.

`event.c` Variable with the number of events in the control group.

`n.c` Variable with the total number of observations in the control group.

`studlab` Variable with the code of the studies.

`META` It accesses the metabin function which allows to modify the arguments of the meta-analysis.

FUNNEL	It accesses the function funnel.meta which allows to modify the arguments of the funnel plot.
LEGEND	It allows to modify the graph legend.
COLOR	It allows to modify the colors of the graph.
file	TXT FILE. Name of the output file with the results.

Details

XII. META-ANALYSIS

XII.4. FUNNEL PLOT

Another problem that can have the meta-analysis (see section of *details* of the function XII1) is the bias in the publication, i.e. the selective publication of the studies on the basis of their results. In other words, that the results that have a specific conclusion be more published and less the studies in which the results give rise to different conclusions.

It has been observed in some studies that if the results are negative, that is to say, that there was no effect, these are typically not published, since it is based on the mistaken idea that a no-effect is not news. It could also be the case that non-publication of the results that were unfavorable to the therapy used with a particular drug. The funnel plot allows to view and assess the bias in the publication on a given topic.

The analysis for the funnel graph comes from the idea that negative studies, which show no effect, are less likely to be published. If there is publication bias, which can occur preferably in small studies in which there is greater likelihood that altering the results by chance, this will tend to publish those that show differences.

In the funnel plot of the differences calculated estimator in the meta-analysis (odds ratio, relative risk, etc.) is represented for each study, along with its standard error. In the case there is no bias, points should be grouped around a central estimator, and would show much greater dispersion around this value smaller outside its size, so that the cloud of points is distributed as an inverted funnel shape.

If it had publication bias in the sense described above, the cloud of points become distorted and the funnel would lose its symmetry. The previous graph shows that the distribution has a funnel shape, confirming the absence of bias that also showed asymmetry test.

FUNCTIONS

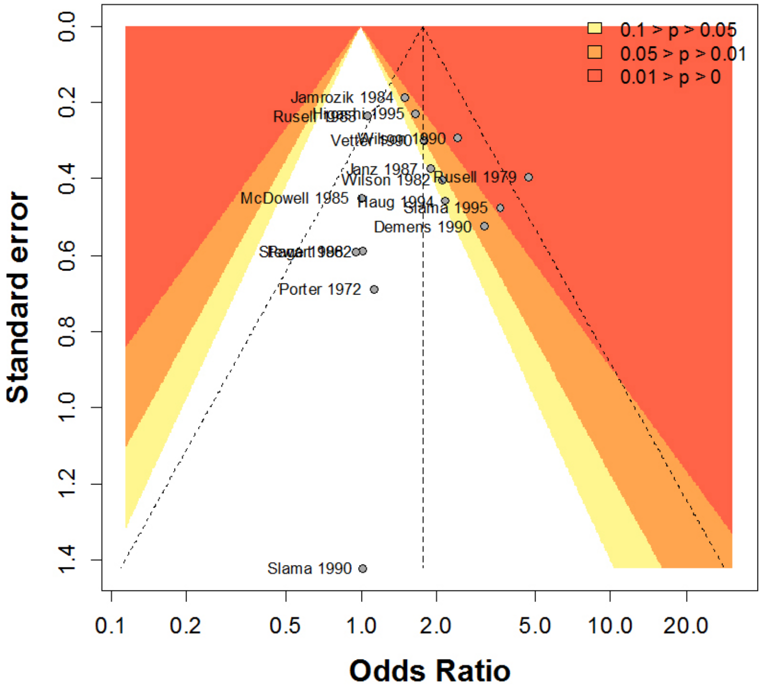
For the meta-analysis the metabin function was used and also for the funnel plot funnel.meta, both from meta package (Schwarzer, 2013). For more details on how to use these functions, from which gives access in META and FUNNEL arguments, consult the help of the function and/or Guisande & Vaamonde (2012).

EXAMPLE

The example is the same as it was explained in detail in the section of *details* of the function XII1.

All the results of the TXT file were also explained in the section of *details* of the function XII1.

Figure XII.4. Funnel plot of a study on
the effect of the doctor's advice to quit smoking.



The graph (Figure XII.4) shown with dotted lines, the funnel to which the data are adjusted, centered on the *OR* estimated. If the points fit reasonably into the funnel, without significant asymmetries, means that there is no evidence of bias (as in the example).

It is also shown, in different shades of colors, the confidence limits (in this example 90%, 95% and 99%), based on the null hypothesis that there is no effect (*OR* = 1) i.e. the funnels-centric value of 1.

The points that are outside of the confidence limits show the studies with significant effects for the corresponding level

Value

A TXT file is obtained with the results of the meta-analysis, in addition to the funnel plot

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Higgins JPT, Thompson SG, Deeks JJ y Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ*: 327-557.

Schwarzer, G. (2013) Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=meta.

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editors. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

Examples

```
## Not run:

data(ZXII1)

#Influence of the medical advice to quit smoking.

XII4(data=ZXII1, event.e=”event.e”, n.e=”n.e”, event.c=”event.c”, n.c=”n.c”, studlab=”Research”,
LEGEND= c(”x=5”, ”y=0.1”,”legend=c('0.1 > p > 0.05','0.05 > p > 0.01', '0 > p > 0.01')”,
”fill=COLOR”, ”bty='n'”), FUNNEL=c(”level=0.95”,
”contour.levels=c(0.9, 0.95, 0.99)”,
”studlab=TRUE”, ”col.contour = COLOR”,
”xlab='Odds Rate (OR)'”, ”ylab='Standard error'”,
”font.lab=2”, ”cex.lab=1.6”, ”cex.axis=1.2”,
”cex.main=1.8”, ”main=''”))


## End(Not run)
```

XII5 *META-ANALYSIS/SENSITIVITY ANALYSIS GRAPH*

Description

A meta-analysis is applied and it is shown in a plot of sensitivity analysis.

Usage

```
XII5(data, event.e, n.e, event.c, n.c, studlab, META=NULL,
FOREST=NULL, file=”Output.txt”)
```

Arguments

`data` Data file.

`event.e` Variable with the number of events in the experimental group.

`n.e` Variable with the total number of observations in the experimental group.

`event.c` Variable with the number of events in the control group.

`n.c` Variable with the total number of observations in the control group.

`studlab` Variable with the code of the studies.

`META` It accesses the metabin function which allows to modify the arguments of the meta-analysis.

`FOREST` It accesses the function forest.meta which allows to modify the arguments of the meta-analysis.

`file` TXT FILE. Name of the output file with the results.

Details

XII. META-ANALYSIS

XII.5. SENSITIVITY ANALYSIS PLOT

In this diagram a sensitivity analysis is performed to determine the influence of each of the studies in the overall estimate of the effect and, therefore, the ruggedness or stability of the final measurement obtained in the meta-analysis.

It is therefore an important complement to the forest plot explained in the section of *details* of the function XII1, within a study of meta-analysis. The method consists in the repetition of the meta-analysis as many times as selected studies, so that every time a study is omitted. If the results of the different meta-analysis are similar, i.e., the effect has the same direction, magnitude and statistical significance, it can be concluded that the result of the meta-analysis is sound.

Sensitivity analysis can also be used to see what happens when the studies which have not been published are eliminated, there are doubts about the quality of the study, etc.

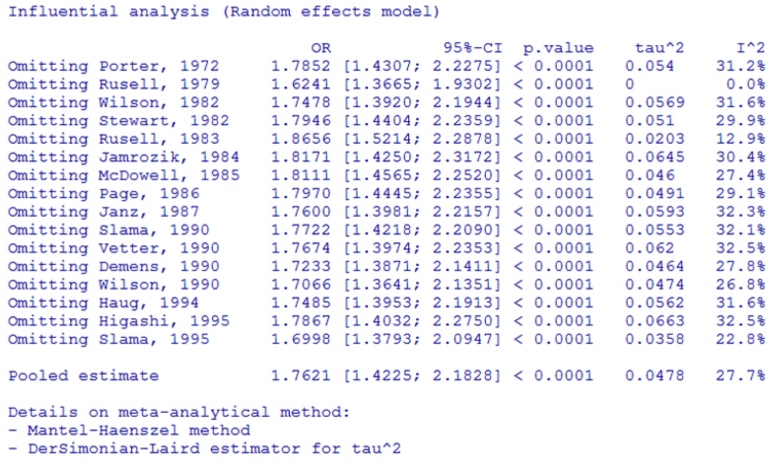
FUNCTIONS

For the meta-analysis the function metabin was used and also for the plot of sensitivity analysis metainf and forest.meta, all from the meta package (Schwarzer, 2013). For more details on how to use these functions, from which gives access in META and FOREST arguments, consult the help of the function and/or Guisande & Vaamonde (2012).

EXAMPLE

The example is the same as it was explained in detail in the section of *details* of the function XII1.

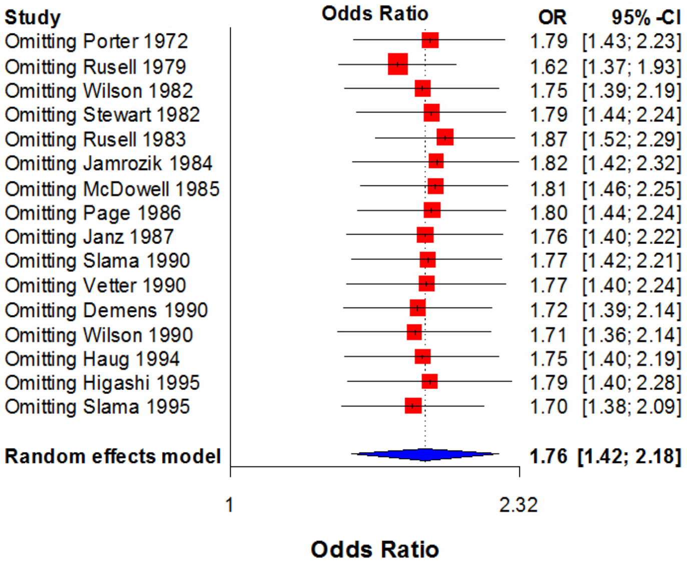
The results -similar to those explained in the previous functions- show the joint OR with its confidence interval obtained by omitting successively each of the studies, the value *P* of the contrast of the aggregate effect, the variance between studies *tau*^2, and the statistical heterogeneity of *I*^2. In this example, the aggregate OR slightly change, major changes occur with the elimination of the studies Russell (1979) and Slama (1995) and in all cases maintaining the heterogeneity in low values, so that none of the studies has a distorting effect on the overall result.



```
Influential analysis (Random effects model)

                            OR           95%-CI  p.value    tau^2      I^2
Omitting Porter, 1972     1.7852 [1.4307; 2.2275] < 0.0001   0.054     31.2%
Omitting Rusell, 1979     1.6241 [1.3665; 1.9302] < 0.0001   0          0.0%
Omitting Wilson, 1982     1.7478 [1.3920; 2.1944] < 0.0001   0.0569    31.6%
Omitting Stewart, 1982    1.7946 [1.4404; 2.2359] < 0.0001   0.051     29.9%
Omitting Rusell, 1983     1.8656 [1.5214; 2.2878] < 0.0001   0.0203    12.9%
Omitting Jamrozik, 1984   1.8171 [1.4250; 2.3172] < 0.0001   0.0645    30.4%
Omitting McDowell, 1985   1.8111 [1.4565; 2.2520] < 0.0001   0.046     27.4%
Omitting Page, 1986       1.7970 [1.4445; 2.2355] < 0.0001   0.0491    29.1%
Omitting Janz, 1987       1.7600 [1.3981; 2.2157] < 0.0001   0.0593    32.3%
Omitting Slama, 1990      1.7722 [1.4218; 2.2090] < 0.0001   0.0553    32.1%
Omitting Vetter, 1990     1.7674 [1.3974; 2.2353] < 0.0001   0.062     32.5%
Omitting Demens, 1990     1.7233 [1.3871; 2.1411] < 0.0001   0.0464    27.8%
Omitting Wilson, 1990     1.7066 [1.3641; 2.1351] < 0.0001   0.0474    26.8%
Omitting Haug, 1994       1.7485 [1.3953; 2.1913] < 0.0001   0.0562    31.6%
Omitting Higashi, 1995    1.7867 [1.4032; 2.2750] < 0.0001   0.0663    32.5%
Omitting Slama, 1995      1.6998 [1.3793; 2.0947] < 0.0001   0.0358    22.8%

Pooled estimate           1.7621 [1.4225; 2.1828] < 0.0001   0.0478    27.7%

Details on meta-analytical method:
- Mantel-Haenszel method
- DerSimonian-Laird estimator for tau^2
```

Figure XII.5. Plot of sensitivity analysis of a study conducted on

the effect of the doctor's advice to stop smoking.



The plot of sensitivity analysis (Figure XII.5) shows that only the work of Russell (1979) has some influence in the analysis, since when it is removed, the overall estimate of the effect changes more than with the other studies.

Value

A TXT file is obtained with the results of the meta-analysis, in addition to the sensitivity analysis plot.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Higgins JPT, Thompson SG, Deeks JJ y Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ*: 327-557.

Schwarzer, G. (2013) Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=meta.

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. En: Lancaster T, Silagy C, Fullerton D. Editors. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

Examples

```
## Not run:
```

```
data(ZXII1)

#Influence of the medical advice to quit smoking.

XII5(data=ZXII1, event.e="event.e", n.e="n.e", event.c="event.c", n.c="n.c", studlab="Research",
FOREST=c("text.random='Random effect model'",
"xlab='Odds ratio (OR)'", "leftlabs=c('Study')",
"col.square='red'","col.diamond='blue'"))



## End(Not run)
```

XIII1 *CORRESPONDENCE ANALYSIS*

Description

An Analysis of Correspondence is applied.

Usage

```
XIII1(data, var, cat=NULL, plot3d=TRUE, ResetPAR=TRUE, PAR=NULL,  xlim=NULL,
ylim=NULL, main=NULL, cex.main=1.7, font.main=2, dim1 = c(1,2), dim2 = c(1,2,3),
map = "symmetric", what = c("all", "all"), mass = c("FALSE", "FALSE"),
contrib = c("none", "none"), col = c("#0000FF", "#FF0000"),
labcol=c("#0000FF", "#FF0000"), pch = c(16, 1, 17, 24), labels = c(2, 2),
arrows = c("FALSE", "FALSE"), LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL,
file="Output.txt")
```

Arguments

`data`	Data file.
`var`	Categorical variables that are included in the analysis.
`cat`	Column with the labels of the cases.
`plot3d`	If it is TRUE, the 3D graph is shown with the three axes selected by the user. This chart does not support accents in the categories of the variable *cat*.
`ResetPAR`	If FALSE, the conditions are not placed by default in the function PAR and those defined by the user on previous graphics are maintained.
`PAR`	It accesses the PAR function which allows to modify many different aspects of the graphic.
`xlim`	Limits of the axis X.
`ylim`	Limits of the axis Y.
`main`	Graph title.
`cex.main`	Font size of the graph title.

font.main Type of font for the title.

dim1 Vector with two values indicating the axis that are represented in the two dimensional graph.

dim2 Vector with three values indicating shafts that are represented in the 3D graph.

map It defines the type of map and this can be any of the following options: "symmetric", "rowprincipal", "colprincipal", "symbiplot", "rowgab", "Colgab", "rowgreen" or "colgreen".

what Vector with two characters that may be "all" or "none", which indicates if the codes are displayed in the cases (first character of the vector) and/or columns (second character of the vector).

mass Vector with two logical values TRUE or FALSE, indicating whether the mass is represented by the area of the symbols, for the cases (first logical value) and/or columns (second logical value).

contrib Vector of two strings of characters that specify if contributions (relative or absolute) should be represented by different color intensities. The available options are "none" (contributions are not listed in the diagram), "absolute" (the absolute contributions are indicated with color intensities) or "relative" (the relative contributions are indicated with color intensities). If "absolute" or "relative" are set, the zero contributions are shown in white. The greater the contribution of a point, the closest to the color defined in the argument *col*.

col Vector with colors for the cases (first color) and columns (second color).

labcol Vector with two colors for the cases and the columns in the 3D graph.

pch Vector with 4 numbers that indicate the symbols, although they are only used the first (for the cases) and the third (for columns).

labels Vector with two values indicating whether the graph should include only symbols (0), only labels (1) or both, symbols and labels (2).

arrows Vector of two logical values that specify if the graph must contain points (FALSE by default) or arrows (TRUE). The first value sets the rows and the second value sets the columns.

LEGEND It allows to embed a legend.

AXIS It allows to add axis.

MTEXT It allows to add text in the margins of the graph.

TEXT It allows to add text in any area of the inner part of the graph.

file TXT FILE. Name of the output file with the results.

Details

XIII. FACTORIAL MULTIVARIATE ANALYSES

XIII.1. ANALYSIS OF CORRESPONDENCE

It is a technique used to represent the rows and columns of a contingency table of two qualitative variables as points in a space of two dimensions, so that they overlap the rows and columns to get a joint representation. In general it is useful in large contingency tables. For example, if we have twenty species of trees and 15 forests and we have done a sample count, we will have a contingency table with the species in rows and columns in the forests. Each box of the table will give us the relative frequency for each dimensional species and forest. The correspondence analysis allows us to represent the species and forests on the same chart, so that it can be associated species with similar distribution profile, forest related, and identify the characteristic species of each forest.

1. Profile of rows and columns

Each row of the table provides a profile row, defined by their frequencies.

1.1. Profile of the *i*-nth row is the proportion of each category of the variable *j* in the row *i*:

$$a_{ij} = \frac{noij}{n_i}$$

The profile of the column totals, i.e., the marginal frequencies of all the columns, define the average profile of the rows.

1.2. Profile of the column *j* -nth is the proportion of each category of the variable *i* in the column *j*:

$$c_{ij} = \frac{noij}{n_j}$$

The profile of the row totals, frequencies of marginal row, is the average profile of the columns.

2. Mass of each element or profile

The mass associated with each row element (r_i) or column (c_j) measures the relative importance of that element in the set, and is equal to its marginal frequency divided by the overall total or sample size *n*, i.e., the corresponding element of the average column profile, while the mass profile of each column will be the corresponding element of the row profile.

$$r_i = \tfrac{n_i}{n} = f_{i.} \quad c_j = \tfrac{n_j}{n} = f_{.j}$$

Representing the row/column profiles, more importance should be given to those with greater mass.

The best measure of distance between profiles (which are defined by the frequency) is not the usual euclidean distance, but the Chi-square distance, since this distance is an important property of distributional equivalence, whereby if two equal profiles are added, the profile added that replaces both maintains the distance of the previous one with any another profile, and will not change any other distance. As the aggregation of rows and columns is usual in the construction of contingency tables, and often subjective criteria are used to decide the number and the limits of the categories or profiles, this property is of special interest. For this reason some simple transformations on the data are carried out to obtain this Chi-square distance.

Chi square distance between row profiles:

$$d^2_{\chi^2}(r_i, r_{i'}) = \sum_{j=1}^{J} \frac{(r_{ij} - r_{i'j})^2}{c_j}$$

Chi square distance between column profiles:

$$d^2_{\chi^2}(c_j, c_{j'}) = \sum_{i=1}^{I} \frac{(c_{ij} - c_{ij'})^2}{r_i}$$

To construct the graph Principal Components method is used, with which generally two dimensions are obtained.

3. Absolute and relative contributions

3.1. Absolute contributions: Part of the variability of each axis or dimension (main component) that is explained by each profile. These serve to interpret the axis according to the profiles, observing which are the profiles that have more weight -or absolute contribution- in the construction of each dimension.

3.2. Part of the variability of each profile explained by each axis or component. They indicate to what extent each profile is adequately explained or represented by the two axes or dimensions. The row or column profiles with small contributions on both axes do not respond to the model, and are poorly represented in the graph, which should be taken into account in its interpretation.

FUNCTIONS

Correspondence analysis is done with the ca function and the graph with the plot.ca function, both from the ca package (Greenacre & Pardo, 2006; Greenacre, 2007; Nenadic & Greenacre, 2007; Greenacre, 2013).

EXAMPLE

Data assessment criteria, on a scale of 0 to 10 of 48 candidates who opt for a job. The objective is to determine if, with a Correspondence Analysis, it is possible to identify which are the qualities that allow to better distinguish between the candidates, in order to facilitate the selection process.

Step 1.

Figures XIII.1 and XIII.2 show that in the right end of the first axis the variables are experience, the suitability for the job and letter of motivation. In the left end of the first axis motivation and confidence are associated. The second axis is associated with the business sense in the upper end, while at the lower end are honesty and level of studies.

Figure XIII.1. 2D Correspondence Analysis showing the variability observed in the qualities of candidates who opt for a job.

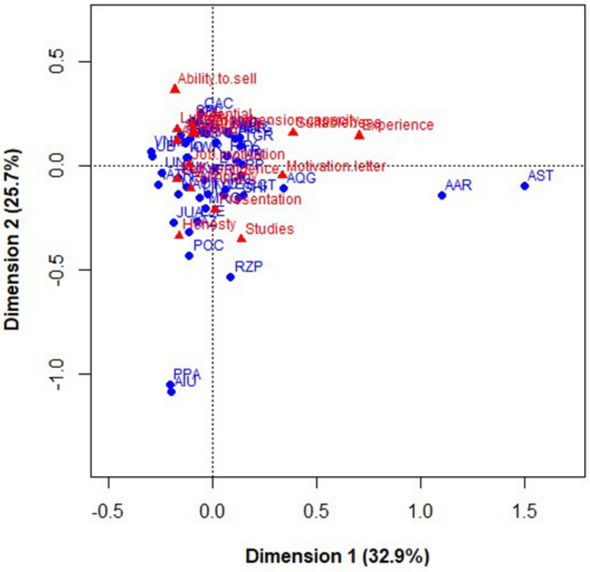
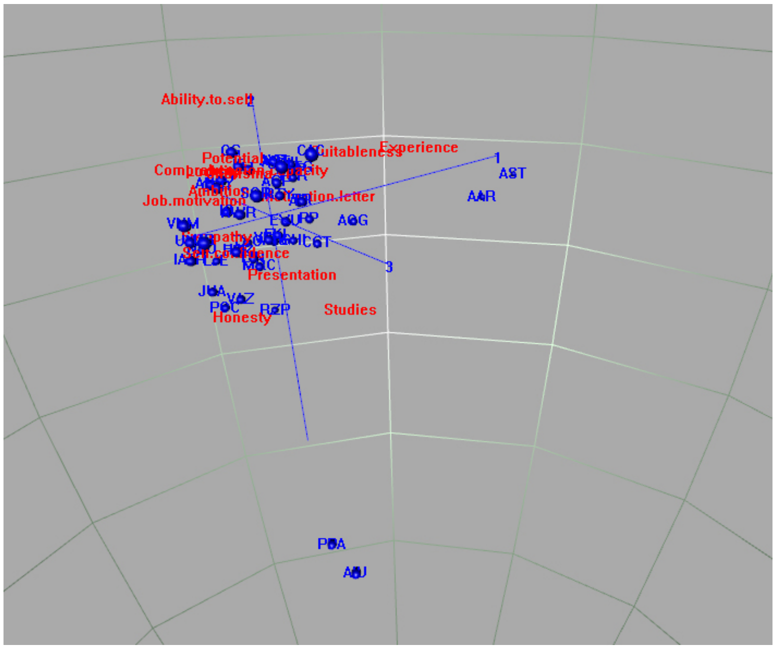
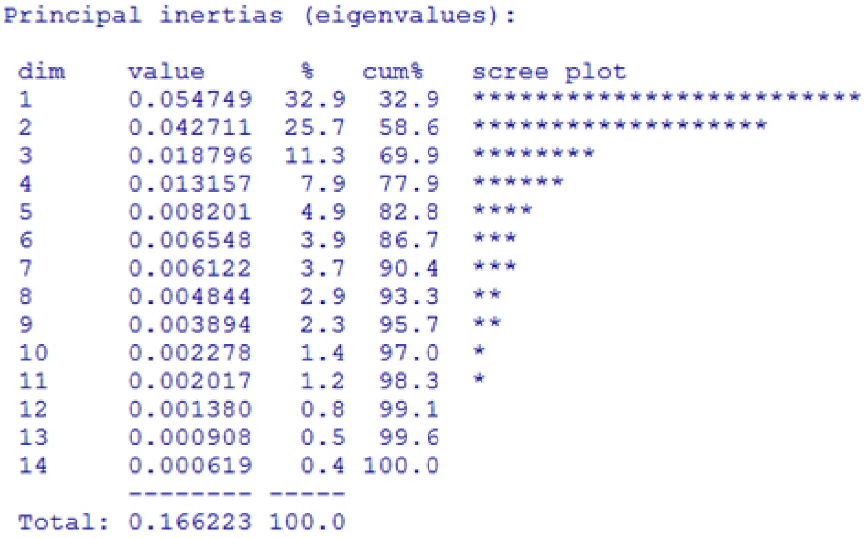


Figure XIII.2. 3D Correspondence Analysis showing the variability observed in the qualities of candidates who opt for a job.

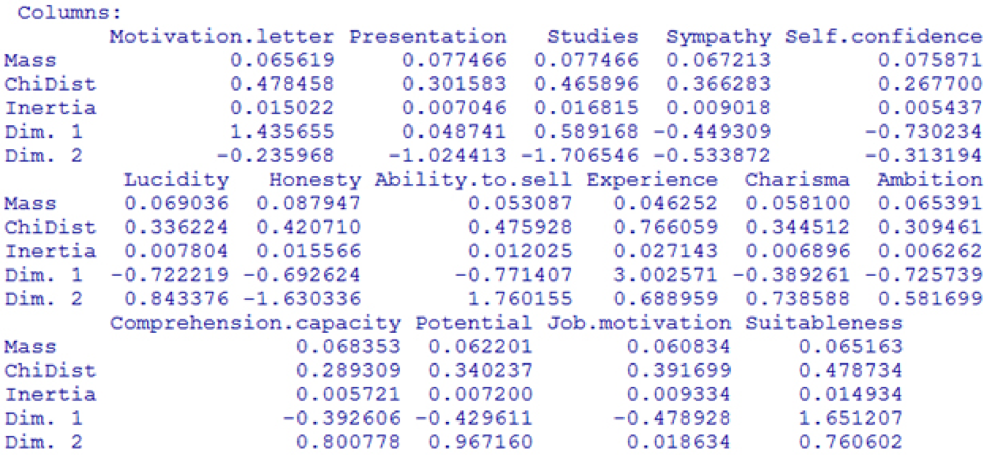


The eigenvalues show the percentage of variance of the axes: 32.9% first axis, 25.7% the second, etc.



```
Principal inertias (eigenvalues):

 dim     value      %    cum%   scree plot
 1       0.054749  32.9   32.9  *************************
 2       0.042711  25.7   58.6  *******************
 3       0.018796  11.3   69.9  ********
 4       0.013157   7.9   77.9  ******
 5       0.008201   4.9   82.8  ****
 6       0.006548   3.9   86.7  ***
 7       0.006122   3.7   90.4  ***
 8       0.004844   2.9   93.3  **
 9       0.003894   2.3   95.7  **
 10      0.002278   1.4   97.0  *
 11      0.002017   1.2   98.3  *
 12      0.001380   0.8   99.1
 13      0.000908   0.5   99.6
 14      0.000619   0.4  100.0
         -------- -----
 Total:  0.166223 100.0
```

Inertia indicates the relative weight or importance of each variable. The highest values are those of the experience (0.0271), educational level (0.0168), honesty (0.0155), letter of motivation (0.015), adequacy (0.0149), commercial sense (0.012, *Ability.to.sell*), etc. The results also show the coordinates of the variables on the axes I (Dim. 1) and II (Dim. 2), which allows to locate each variable in the graph shown above in order to dispose the coordinates to perform any other type of representation.

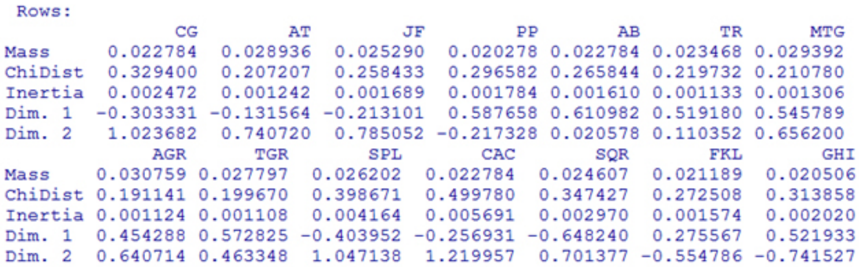


```
 Columns:
        Motivation.letter Presentation   Studies  Sympathy Self.confidence
Mass             0.065619     0.077466  0.077466  0.067213        0.075871
ChiDist          0.478458     0.301583  0.465896  0.366283        0.267700
Inertia          0.015022     0.007046  0.016815  0.009018        0.005437
Dim. 1           1.435655     0.048741  0.589168 -0.449309       -0.730234
Dim. 2          -0.235968    -1.024413 -1.706546 -0.533872       -0.313194
         Lucidity   Honesty Ability.to.sell Experience  Charisma  Ambition
Mass     0.069036  0.087947        0.053087   0.046252  0.058100  0.065391
ChiDist  0.336224  0.420710        0.475928   0.766059  0.344512  0.309461
Inertia  0.007804  0.015566        0.012025   0.027143  0.006896  0.006262
Dim. 1  -0.722219 -0.692624       -0.771407   3.002571 -0.389261 -0.725739
Dim. 2   0.843376 -1.630336        1.760155   0.688959  0.738588  0.581699
        Comprehension.capacity Potential Job.motivation Suitableness
Mass                  0.068353  0.062201       0.060834     0.065163
ChiDist               0.289309  0.340237       0.391699     0.478734
Inertia               0.005721  0.007200       0.009334     0.014934
Dim. 1               -0.392606 -0.429611      -0.478928     1.651207
Dim. 2                0.800778  0.967160       0.018634     0.760602
```

The most important instrument in the correspondence analysis is a graph in which the row and column objects related by distances and vicinity which can be interpreted. Among the results the (*ChiDist*) distances are shown to average profile.

The mass associated with each row or column element -similar and alternative to the concept of inertia- measures the importance of that element in the set, and is equal to its marginal relative frequency. Representing the row-column profiles greater importance should be given in interpretation to those with greater mass.

The inertia of the candidates, which will choose the best candidate based on the quality to which priority is also shown, as well as the coordinate at the axes I (Dim. 1) and II (Dim. 2).



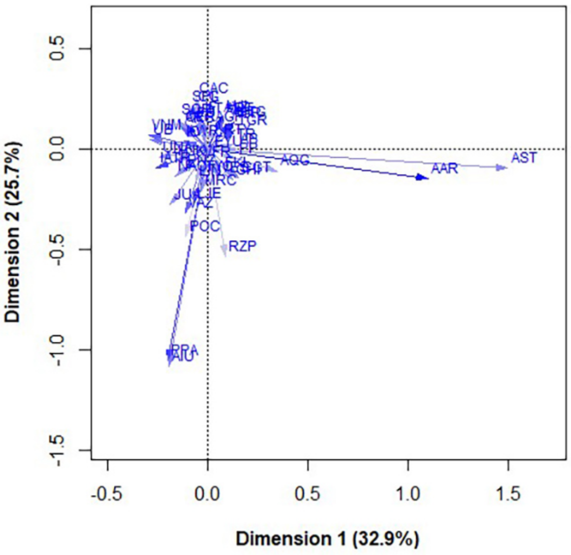
```
 Rows:
               CG        AT        JF        PP       AB       TR      MTG
Mass     0.022784  0.028936  0.025290  0.020278 0.022784 0.023468 0.029392
ChiDist  0.329400  0.207207  0.258433  0.296582 0.265844 0.219732 0.210780
Inertia  0.002472  0.001242  0.001689  0.001784 0.001610 0.001133 0.001306
Dim. 1  -0.303331 -0.131564 -0.213101  0.587658 0.610982 0.519180 0.545789
Dim. 2   1.023682  0.740720  0.785052 -0.217328 0.020578 0.110352 0.656200
             AGR      TGR       SPL       CAC       SQR       FKL       GHI
Mass    0.030759 0.027797  0.026202  0.022784  0.024607  0.021189  0.020506
ChiDist 0.191141 0.199670  0.398671  0.499780  0.347427  0.272508  0.313858
Inertia 0.001124 0.001108  0.004164  0.005691  0.002970  0.001574  0.002020
Dim. 1  0.454288 0.572825 -0.403952 -0.256931 -0.648240  0.275567  0.521933
Dim. 2  0.640714 0.463348  1.047138  1.219957  0.701377 -0.554786 -0.741527
```

Step 2.

It performs the Correspondence Analysis showing only the candidates (cases), which is shown in Figure XIII.3.

Figure XIII.3. Correspondence Analysis showing the variability observed

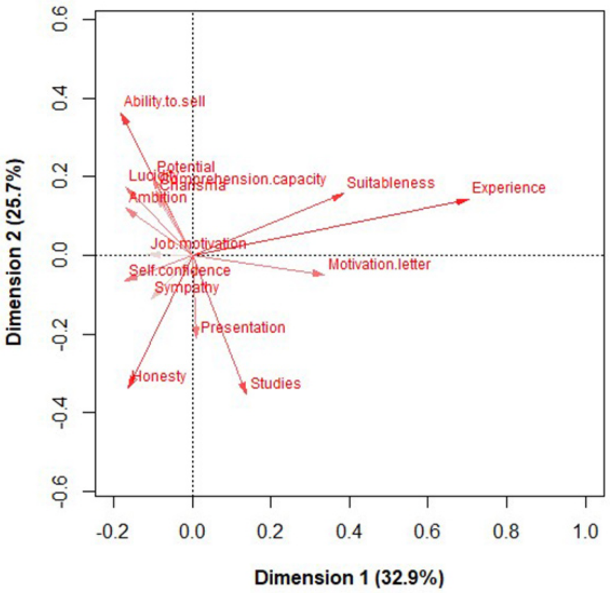
in the qualities of candidates who opt for a job.



Step 3.

It performs the Correspondence Analysis showing only the qualities of the candidates (columns), which is shown in the Figure XIII.4.

Figure XIII.4. Correspondence Analysis showing the variability observed in the qualities.



If to select the candidate we rely on the variables that most influenced the first axis in its rightmost (experience and adequacy), candidates who have the most positive coordinates on the first axis are AST and AAR. If on the other hand for the selection we rely on the variables that best discriminated against the second axis, as for example business sense which was located in the upper end of the positive axis, so the selected candidate should be ACC. If we chose the honesty, the selected candidate should be PPA or AIU.

Value

A TXT file is obtained with the results of the Correspondence Analysis, a graph with scores and the 3D graph if this is selected.

References

Nenadic, O. & Greenacre, M. (2007) Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20: 1-13

Greenacre, M. (2013). Simple, Multiple and Joint Correspondence Analysis. R package version 0.53. Available at: `http://CRAN.R-project.org/package=ca`.

Greenacre, M. (2007) *Correspondence Analysis in Practice*. Second Edition. London: Chapman & Hall / CRC.

Greenacre, M.J. & Pardo, R. (2006) Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods and Research*, 35: 193-218.

Examples

```
## Not run:
data(ZXIII1)

#Step 1
XIII1(data=ZXIII1, cat="Candidate", var=c("Motivation.letter","Presentation","Studies",
"Sympathy","Self.confidence", "Lucidity","Honesty","Ability.to.sell","Experience",
"Charisma","Ambition","Comprehension.capacity","Potential","Job.motivation",
"Suitableness"), xlim=c(-0.5,1.7))

#Step 2
XIII1(data=ZXIII1, cat="Candidate", var=c("Motivation.letter","Presentation","Studies",
"Sympathy","Self.confidence", "Lucidity","Honesty","Ability.to.sell","Experience",
"Charisma","Ambition", "Comprehension.capacity","Potential","Job.motivation",
"Suitableness"), what = c("all", "none"), mass = c("TRUE","FALSE"), contrib = "relative",
arrows = c("TRUE", "FALSE") , xlim=c(-0.5,1.7))

#Step 3
XIII1(data=ZXIII1, cat="Candidate", var=c("Motivation.letter","Presentation","Studies",
"Sympathy", "Self.confidence", "Lucidity","Honesty","Ability.to.sell", "Experience",
"Charisma","Ambition","Comprehension.capacity","Potential","Job.motivation",
"Suitableness"), xlim=c(-0.2, 1), what = c("none", "all"), mass = c("FALSE","TRUE"),
contrib = "relative",  arrows = c("FALSE", "TRUE"))

## End(Not run)
```

XIII2 *PRINCIPAL COMPONENT ANALYSIS*

Description

An Principal Component Analysis is applied.

Usage

```
XIII2(data, var, cat=NULL, labels=NULL, ellipse=FALSE, convex=FALSE, dim=c(1,2),
VIF=FALSE, threshold=10, ResetPAR=TRUE, PAR=NULL, PCA=NULL, SCATTERPLOT=NULL,
COLOR=NULL, PCH=NULL, XLIM=NULL, YLIM=NULL, XLAB=NULL, YLAB=NULL, LEGEND=NULL,
AXIS=NULL, MTEXT= NULL, TEXTvar=NULL, TEXTlabels=NULL, arrows=TRUE, larrow=0.7,
colArrows="black", file1="Output.txt", file2="Var loadings.csv",
file3="Cat loadings.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

`data`	Data file.
`var`	Variables that are included in the analysis.
`cat`	Categories of cases.
`labels`	Variable that allows to display a label for each case.

`ellipse` If it is TRUE, the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *cat* are depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*.

`convex` If it is TRUE the convex hull is depicted for each category.

`dim` Vector with two values indicating the axes that are shown in the graph.

`VIF` If it is TRUE, the inflation factor of the variance (VIF) is used to select the highly correlated variables and, therefore, not correlated variables are excluded from the analysis.

`threshold` Cut-off value for the VIF.

`ResetPAR` If it is FALSE, the default condition of the function PAR are not placed and those defined by the user on previous graphics are maintained.

`PAR` It accesses thePAR function that allows to modify many different aspects of the graphs.

`PCA` It accesses the prcomp function of the stats package.

`SCATTERPLOT` It accesses the function scatterplot of the car package.

`COLOR` It allows to modify the colors of the graph, but they must be as many as different groups the variable *cat* has.

`PCH` Vector with the symbols on the chart, which must be as many as different groups the variable *cat* has. If it is NULL, they are automatically calculated starting with the symbol 15.

`XLIM, YLIM` Vectors with the limits of the axes *X* and *Y*.

`XLAB, YLAB` Legends of the axes *X* and *Y*.

`LEGEND` It allows to include or to modify a legend.

`AXIS` It allows to add axes.

`MTEXT` It allows to add text in the margins of the graph.

`TEXTvar` It allows to modify the labels of the variables in the graph.

`TEXTlabels` It allows to modify the labels of the cases in the graph.

`arrows` If it is TRUE the arrows are shown in the scatterplot.

`larrow` It modifies the length of the arrows.

`colArrows` Colors of the arrows.

`file1` TXT FILE. Name of the output file with the results.

`file2` CSV FILE. Name of the output file with the coordinates of the variables in the graph.

`file3` CSV FILE. Name of the output file with the coordinates of the cases in the graph.

`na` CSV FILES. Text that is used in the cells without data.

`dec` CSV FILES. It defines if a comma " ," or a dot " ." is used as decimal separator.

`row.names` CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

XIII. FACTORIAL MULTIVARIATE ANALYSES

XIII.2. PRINCIPAL COMPONENT ANALYSIS

It consists of the transformation of the whole original variables in another set of variables -Principal Components- obtained as a linear combination of those. The new variables in the same number as the old ones still retain all the information -in what refers to the variability of the set of objects- of the primitive variables, but most of the major components have a variability so small that can be ignored, in such a way that a few components -usually three or fewer- allow to represent and reasonably explain the set of objects in the sample without significant loss of information.

The main virtue of the Principal Component Analysis (PCA) consists precisely in the reduction of the complexity of the data, the passing of many variables to few (often one or two), that can be represented graphically. Against this advantage, the drawback of the new variables is in the partial loss of information (often small) and in the fact that the original variables have a real meaning of lacking basic components, which are a mixture or combination of variables.

The first main component is a linear combination of the original variables that has maximum variance. The second component is the linear combination of the original variables with maximum variance with the added condition to become independent from the first (orthogonal), and thus, all the main components can be obtained, which remaining independent among them, contain different information, since the independence or absence of correlation means that the new variables or components do not share common information. Each principal component explains therefore the maximum possible residual variability (that have not already explained the previous ones).

To find the optimal combinations of the original variables, the eigenvalues and eigenvectors of the matrix communality covariance (or correlation matrix) should be calculated. The eigenvalues, sorted from highest to lowest are the variances of the new components, and eigenvectors express the linear combinations that define the major components.

It should be decided how many components must be preserved: a high number allows to explain a larger proportion of the total variability, and a low number allows to obtain for greater simplicity in the graphical representation. There are several criteria that can help when it comes to rule:

- Criterion of the variance. The components explaining a sufficient percentage (e.g. 80% or 90%) variance are selected.

- Criterion of Kaiser. The components whose variances (eigenvalues) are higher than the average variance are preserved.

- Graph of sedimentation (Scree plot). The eigenvalues that generally decrease rapidly at the beginning and slowly after are represented, and only those that are above the elbow of the curve are preserved.

- Two components. Keeping only the first two allows to build a graph in two dimensions which is usually the most appropriate to describe the dataset.

The only condition for the Principal Component Analysis can be applied is that variables be quantitative. However, there is another condition or assumptions necessary to make its application to be useful: the original variables should be correlated. The higher the correlation between the variables, the greater the proportion of variance explained by the first components and, therefore, the smaller the loss of information to keep them apart from the rest.

To test if the variables are correlated, there are several procedures:

- Inspect the matrix of correlations. It could be observed if the majority of the coefficients of linear correlation between the original variables are close to 1 or -1.

- Inspect the matrix of values of *p* of the contrast of the invalidity of correlation coefficients. It contrasts -for each one of the coefficients of correlation- that the hypothesis is zero. The value *p* must be less than the significance level (usually 0.05) to reject this hypothesis. In this way, the matrix of values of *p*, when there is correlation, must be comprised mostly of zeros or very small values.

- Determinant of the matrix of correlations. Takes a value close to 1 when there is no correlation and next to zero when there is correlation.

- Test of sphericity Bartlett. It is based on the determinant of the matrix of correlations and tests the hypothesis of "sphericity" or absence of correlation with a single test. A value *p* of the contrast lower than the level of chosen significance allows to reject the hypothesis and conclude that there is a correlation.

When the original variables are heterogeneous, that is to say, are expressed in different units of measure, it happens that the variability is determined by the choice of the unit of measure for each of them, which influences the results of the implementation of the ACP when assigning an arbitrary importance to each variable. To avoid this negative influence, the solution is to use the matrix of correlations in place of the covariance matrix, which is equivalent to typify the original variables. In this way all the variables will have average 0 and standard deviation 1. As this process involves a certain loss of information, in the case of homogeneous variables, the matrix of correlations in replacement of the covariance matrix should not be used.

Once obtained the linear combinations that define the major components, we can interpret the coefficients (saturations or loads): the higher its absolute value, has greater influence that variable in the construction of the component. It is possible to interpret the components, and assign generic meanings, from the set of variables with the greatest weight in its construction.

To facilitate this interpretation of the components and assign them any meaning, the graph of saturations can be useful, because that shows -at the level of the first two components- the coefficients or saturations for each variable, so that each original variable is represented as a point in this plane. The variables that are more remote with respect to the zero of the horizontal axis are the ones that give meaning to the first component (that often expresses a generic idea of "size"), and those that are more distant relative to the zero of the vertical axis are those that give meaning to the second component (often reflects a generic idea of "shape"). The variables near the center of coordinates have no interest, and the variables that are moving away from both axes are common to both components and not helping their interpretation. The graph can also be made for other components different from the first two.

The main result of this technique is in general the chart of the scores. This is the representation, in the plane of the first two components, the coordinates of the points that represent the elements of the set object (rows or cases in our data). This representation allows to describe the set of multidimensional data in a relatively easy way. It may be useful to add the chart case labels or brands. The original variables have to be correlated to make the method work well. The use of the inflation factor of the variance (VIF) that is proposed as a criterion for optionally select the variables that must be used in the ACP, allows to exclude those that are poorly correlated with the remaining ones, which in general would appear associated with unimportant principal components (other than the first), but this is not necessarily always the case, for what this automatic selection of variables should be used with caution, only as an aid in the selection of variables to be used.

FUNCTIONS

The Principal Components Analysis was performed with the prcomp function of the stats package. The vif function of the usdm package was used for the calculation of VIF (Naimi, 2013; Naimi et al., 2014). To perform the *biplot* graph the scatterplot function of the car package was used (Fox et al., 2014). The arrows are depicted with the function Arrows of the package IDPmisc (Locher & Ruckstuhl, 2014). The convex hull is estimated with the function chull of the package grDevices. KMO test was performed with the function KMO of the package psych (Revelle, 2018). Bartlett's test sphericity was performed with the function bart_spher of the package REdaS (Maier, 2015).

EXAMPLE

The study was to analyze the demographic parameters of 57 countries in Europe, Africa and America. The variables used were male and female life expectancy at birth (in years of life), the mortality rates, infant mortality, birth, and fertility, the gross domestic product per capita (in thousands of dollars per year) and the literacy rate for men and women (in percentage) in the year 2000. The data were obtained from The World Bank (`http://www.worldbank.org/`). The objective is to determine which are the demographic parameters to show greater variability among countries and, therefore, which are responsible for the differences that exist among them.

Step 1.

First, all variables are used *var=c("LifeExpF", "LifeExpM", "Mortinf", "PIB_cap", "Birthrate", "Mortality", "Fertility", "LiteracyM", "LiteracyF")* and as categories or classification variable continent is used *cat="Continent"*. Then the analysis is performed by eliminating the variables that are uncorrelated, for what it is necessary to specify *VIF=TRUE*, which means that it will eliminate those with VIF below the selected threshold, which by default is *threshold= 10*. For this reason, mortality, and the GDP will be deleted automatically.

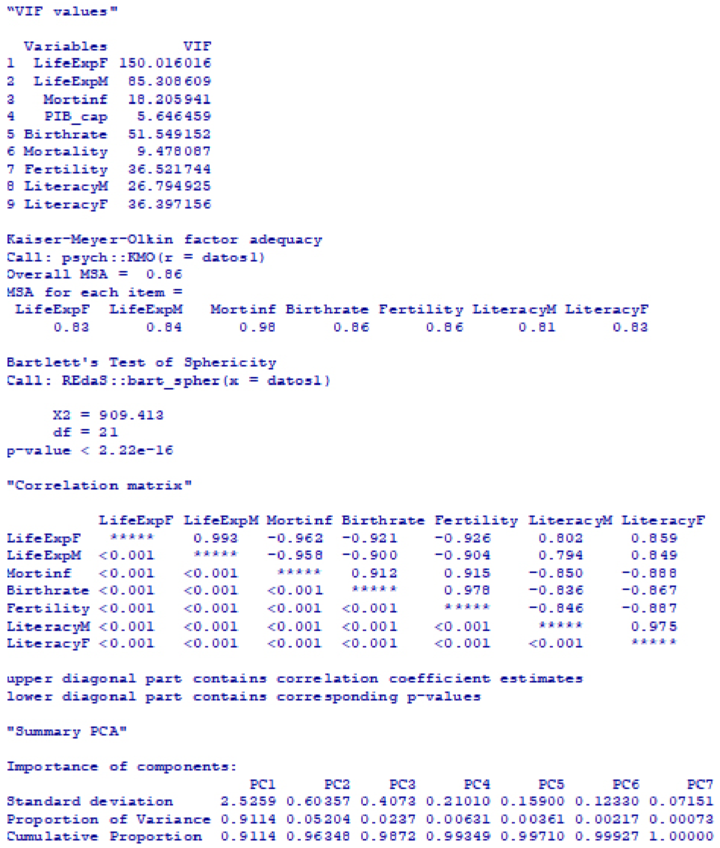
The first thing that is displayed in the results are the values of VIF. As mentioned above, the original variables must be correlated, to achieve a greater proportion of variance explained by the first components.

The next result is the KMO test, which tells us if the variables are suitable for the Principal Components Analysis. The value must be greater than 0.5. Therefore, all variables that do not have a value greater than 0.5, could be eliminated from the analysis. In the case that the value is exactly 0.5, it means that is not possible to estimate the KMO.

The second statistic is Bartlett's sphericity test. A value *p* of the contrast smaller than 0.05 allows rejecting the hypothesis and concluding that there is correlation. Therefore, for the Principal Components Analysis to be valid, the probability must be less than 0.05, as it is in this case.

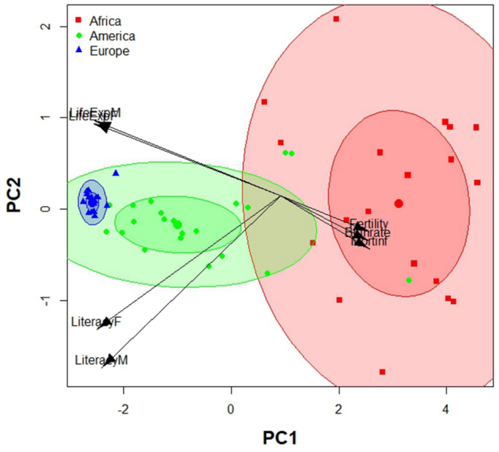
As can be seen in the matrix of values of *p*, all are less than 0.05 which indicates that none of the coefficients of correlation seems to be null. Therefore, all variables are correlated. The results also show the relative contribution of each axis.

The first axis explains 91.1%, the second 5.2% and the third 0.2% of the observed variation. The first two axes explain 96.3% of the variance. In Figure XIII.5 is observed that in the first axis, the countries of Europe, America and Africa are well differentiated. Those in Europe are characterized by high life expectancy and high literacy, in both men and women. The countries of Africa due to its high infant mortality and high birth rate. The ellipses show the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each of the sampling sites. It is an indicator of the degree of overlap between the continents. These levels can be modified by entering in the scatterplot function using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*. With some examples it is observed that it is not possible to estimate the ellipses and this error *Error in svd(sqrt(w/sum(w)) * X, nu = 0):* is obtained or this other *Error in cov.trob(cbind(x[use], y[use]),* *wt = weights[use]): no positive weights*. In case of obtaining these errors or simply not wanting to show the ellipses, again using the argument *ellipse=FALSE*.



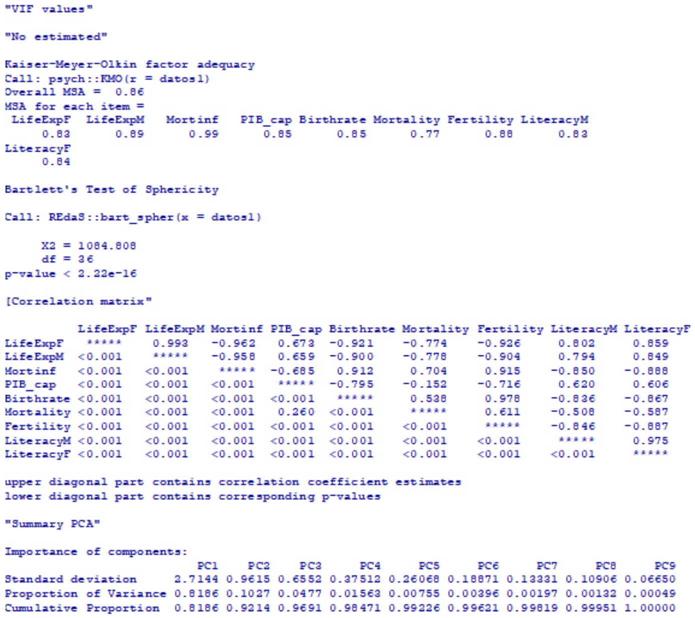
```
"VIF values"

  Variables        VIF
1  LifeExpF 150.016016
2  LifeExpM  85.308609
3   Mortinf  18.205941
4   PIB_cap   5.646459
5 Birthrate  51.549152
6 Mortality   9.478087
7 Fertility  36.521744
8 LiteracyM  26.794925
9 LiteracyF  36.397156

Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = datos1)
Overall MSA =  0.86
MSA for each item =
 LifeExpF  LifeExpM   Mortinf Birthrate Fertility LiteracyM LiteracyF
     0.83      0.84      0.98      0.86      0.86      0.81      0.83

Bartlett's Test of Sphericity
Call: REdaS::bart_spher(x = datos1)

     X2 = 909.413
     df = 21
p-value < 2.22e-16

"Correlation matrix"

          LifeExpF LifeExpM Mortinf Birthrate Fertility LiteracyM LiteracyF
LifeExpF   *****    0.993   -0.962  -0.921    -0.926     0.802     0.859
LifeExpM  <0.001    *****   -0.958  -0.900    -0.904     0.794     0.849
Mortinf   <0.001   <0.001    *****   0.912     0.915    -0.850    -0.888
Birthrate <0.001   <0.001   <0.001   *****     0.978    -0.836    -0.867
Fertility <0.001   <0.001   <0.001  <0.001     *****    -0.846    -0.887
LiteracyM <0.001   <0.001   <0.001  <0.001    <0.001     *****     0.975
LiteracyF <0.001   <0.001   <0.001  <0.001    <0.001    <0.001     *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values

"Summary PCA"

Importance of components:
                          PC1     PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5259 0.60357 0.4073 0.21010 0.15900 0.12330 0.07151
Proportion of Variance 0.9114 0.05204 0.0237 0.00631 0.00361 0.00217 0.00073
Cumulative Proportion  0.9114 0.96348 0.9872 0.99349 0.99710 0.99927 1.00000
```

Figure XIII.5. 2D Principal Component Analysis showing the observed variability in demographic parameters of the 57 analyzed countries.



Step 2.

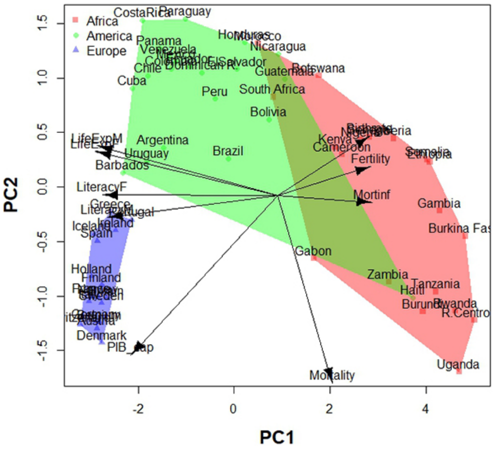
The analysis is done but without deleting any variable, so the option by default *VIF=FALSE* is left. With *labels="Country"*, the cases that are the names of the countries are specified. Convex hull is used instead of the ellipses with the arguments *convex=TRUE* and *ellipse=FALSE*.



```
"VIF values"

"No estimated"

Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = datos1)
Overall MSA =  0.86
MSA for each item =
 LifeExpF  LifeExpM   Mortinf   PIB_cap Birthrate Mortality Fertility LiteracyM
     0.83      0.89      0.99      0.85      0.85      0.77      0.88      0.83
LiteracyF
     0.84

Bartlett's Test of Sphericity

Call: REdaS::bart_spher(x = datos1)

     X2 = 1084.808
     df = 36
p-value < 2.22e-16

[Correlation matrix"

          LifeExpF LifeExpM Mortinf PIB_cap Birthrate Mortality Fertility LiteracyM LiteracyF
LifeExpF   *****    0.993   -0.962   0.673  -0.921    -0.774    -0.926     0.802     0.859
LifeExpM  <0.001    *****   -0.958   0.659  -0.900    -0.778    -0.904     0.794     0.849
Mortinf   <0.001   <0.001    *****  -0.685   0.912     0.704     0.915    -0.850    -0.888
PIB_cap   <0.001   <0.001   <0.001   *****  -0.795    -0.152    -0.716     0.620     0.606
Birthrate <0.001   <0.001   <0.001  <0.001   *****     0.538     0.978    -0.836    -0.867
Mortality <0.001   <0.001   <0.001   0.260  <0.001     *****     0.611    -0.508    -0.587
Fertility <0.001   <0.001   <0.001  <0.001  <0.001    <0.001     *****    -0.846    -0.887
LiteracyM <0.001   <0.001   <0.001  <0.001  <0.001    <0.001    <0.001     *****     0.975
LiteracyF <0.001   <0.001   <0.001  <0.001  <0.001    <0.001    <0.001    <0.001     *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values

"Summary PCA"

Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.7144 0.9615 0.6552 0.37512 0.26068 0.18871 0.13331 0.10906 0.06650
Proportion of Variance 0.8186 0.1027 0.0477 0.01563 0.00755 0.00396 0.00197 0.00132 0.00049
Cumulative Proportion  0.8186 0.9214 0.9691 0.98471 0.99226 0.99621 0.99819 0.99951 1.00000
```

As expected, there are values of p>0.05 in the correlation matrix, which means that not all variables are correlated. In addition, it is also observed that the first axis explains 81.86% of the observed variability, the second 10.2% and the third the 4.77%. The two first axes explained 92.1% of the variance, which is lower, as was expected, where only the correlated variables (96.3%) were used. Again continents (Figure XIII.7) are separated, where Europe is distinguished by a higher GDP.

Figure XIII.7. 2D Principal Component Analysis showing the observed variability in demographicparameters, without removing uncorrelated ones, of the 57 analyzed countries.



Value

A TXT file with the VIF is obtained, the correlations between variables, and the results of the Analysis of Principal Components. In addition a 2D graph representing the variables and categories is obtained.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: http://CRAN.R-project.org/package=IDPmisc.

Maier, M.J. (2015) Companion Package to the Book 'R: Einführung durch angewandte Statistik. R package version 0.9.3. Available at: http://CRAN.R-project.org/package=REdaS.

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Revelle,W. (2018) Procedures for Psychological, Psychometric, and Personality Research. R package version 1.8.4. Available at: http://CRAN.R-project.org/package=psych.

Examples

```
## Not run:
data(ZXIII2)
```

```
#Step 1
XIII2(data=ZXIII2, var=c("LifeExpF", "LifeExpM", "Mortinf", "PIB_cap", "Birthrate",
"Mortality", "Fertility", "LiteracyM", "LiteracyF"), cat="Continent",
VIF=TRUE, ellipse=TRUE)

#Step 2
XIII2(data=ZXIII2, var=c("LifeExpF", "LifeExpM", "Mortinf",
"PIB_cap", "Birthrate",
"Mortality", "Fertility", "LiteracyM", "LiteracyF"), cat="Continent",
labels="Country", convex=TRUE, XLIM=c(-4,6), YLIM=c(-2,1.7))

## End(Not run)
```

XIII3 *MULTIDIMENSIONAL SCALING*

Description

A multidimensional scaling is applied.

Usage

```
XIII3(data, var, cat=NULL, ellipse=FALSE, convex=FALSE, dim=c(1,2), ResetPAR=TRUE,
PAR=NULL, METAMDS=NULL, SCATTERPLOT=NULL, COLOR=NULL, PCH=NULL, XLIM=NULL,
YLIM=NULL, XLAB=NULL, YLAB=NULL, TEXT=NULL, LEGEND=NULL, AXIS=NULL, MTEXT=NULL,
arrows=TRUE, larrow=0.95, colArrows="black", file1="Var loadings.csv",
file2="Cat loadings.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

`data`	Data file.
`var`	Variables included in the analysis.
`cat`	Categories of the cases.
`ellipse`	If it is TRUE, the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *cat* are depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*.
`convex`	If it is TRUE the convex hull is depicted for each category.
`dim`	Dimensions that are represented on the graph.
`ResetPAR`	If it is FALSE, the default condition of the function PAR are not placed and those defined by the user on previous graphics are maintained.
`PAR`	It accesses the function PAR that allows to modify different aspects of the graph.
`METAMDS`	It accesses the function metaMDS from the vegan package.
`SCATTERPLOT`	It accesses the function scatterplot from the car package.
`COLOR`	It allows to modify the colors of the graph, but must be as many different groups as the variable has *cat*.

`PCH` Vector with the graph symbols, which must be as many different groups as the variable has *cat*. If it is NULL, this is automatically calculated starting with the symbol 15.

`XLIM, YLIM` Vectors with the limits of the axes *X* and *Y*.

`XLAB, YLAB` Legends of the axes *X* and *Y*.

`TEXT` It accesses the function which allows to change the labels of cases in the graph *biplot*.

`LEGEND` It includes a legend.

`AXIS` It adds axes.

`MTEXT` It allows to add text in the margins of the graph.

`arrows` If it is TRUE the arrows are shown in the scatterplot.

`larrow` It modifies the length of the arrows.

`colArrows` Color of the arrows.

`file1` CSV FILE. Name of the output file with the coordinates of the variables in the graph.

`file2` CSV FILE. Name of the output file with the coordinates of the rows or cases, adding categories in the graph.

`na` CSV FILE. Text used in the cells without data.

`dec` CSV FILE. It defines if a comma"," o a dot "." can be used as decimal separator.

`row.names` CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

XIII. FACTORIAL MULTIVARIATE ANALYSIS

XIII.3. MULTIDIMENSIONAL SCALING

It's about finding the underlying structure of a set of measures of proximity/distance between objects. Objects can be of any type: Countries, trademarks, kinship relationships, businesses, descriptors of quality, etc., the measures that allow to set proximity or distances between objects may also be of different types, quantitative or qualitative, real or supposed, objective or subjective. This technique is typically applied to the study of perceptions, of great interest in psychology, marketing, quality control, and other areas.

If we have a map, it is relatively simple to build a table of distances between towns. The multidimensional scaling is trying to solve the inverse problem: from a table of distances, a map is constructed, which allows to place objects and graphically interpret the relationships between them.

Each observation is assigned to a specific position in a space of small size, usually two, in such a way that the distances between the points correspond the best possible way with the distances between objects. The main result of this technique is therefore a graph.

The criterion or extent of most important adjustment used in the multidimensional scaling is the Stress (Kruskal), which is an average of deviations between end distances on the map and the distances or initial dissimilarities, normalized to take values between 0 and 1.

$$S = \frac{\sum\sum\left(d_{ij} - \bar{d}_{ij}\right)^2}{\sum\sum d_{ij}^2}$$

The greater the stress, the worse the fit. It is generally recommended not to use this technique when a value is greater than 0.2. It is often expressed as a percentage: the fit is good when the stress is less than 5% (0.05). For its interpretation the following indicative scale can be used:

Stress	**Adjustment**
0,1-0,2	Poor
0,05-0,1	Regular
0,025-0,05	Good
<0,025	Excellent
0	Perfect

The squared multiple correlation coefficient RSQ between gaps and disparities, which can be interpreted as the proportion of variance of the transformed data -or percentage of variability- collected by the obtained representation is also used as a setting. It varies between 0 and 1, and the closer to 1 the better the fit; 0.6 is generally considered a minimum acceptable value. There are numerous variants and algorithms for the realization of the multidimensional scaling.

FUNCTIONS

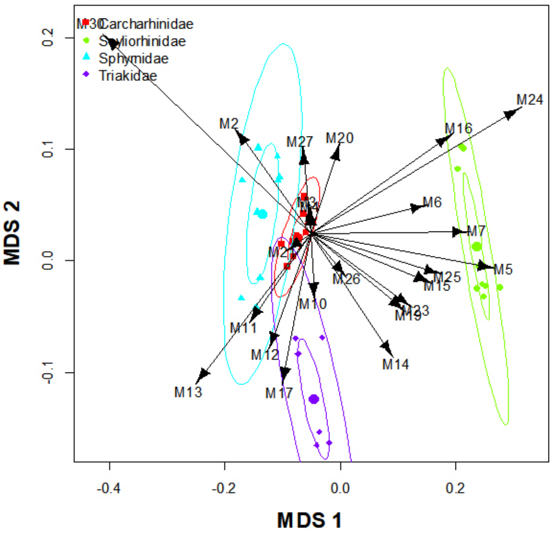
The Multidimensional Scaling was performed with the metaMDS of the vegan package (Oksanen et al., 2013). To make the *biplot*graph, the scatterplot function of the car package (Fox et al., 2014) was used. The arrows are depicted with the function Arrows of the package IDPmisc (Locher & Ruckstuhl, 2014). The convex hull is estimated with the function chull of the package grDevices.

EXAMPLE

The study consisted of performing different morphometric measures of shark species, belonging to different families of the order Carcharhiniformes. The aim is to determine if species of the same family have similar morphometrics and, therefore, the Multidimensional Scaling shows that there is a morphometric pattern compared to each of the families.

In the results window the value of the measure of adjustment is shown, with a stress of 0.009 (0.9%), which according to the classification presented in the introduction is considered good. It is noted that there is a relatively good clustering between families (Figure XIII.8). Therefore, it is concluded that it is possible to differentiate the different families of sharks in function of its morphometry. The *graph* biplot also allows to see which are the morphometric variables responsible for this differentiation between families. The ellipses show the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each of the families. It is an indicator of the degree of overlap between the families of sharks. These levels can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*.

Figure XIII.8. Multidimensional scaling using morphometric measures from different families of sharks.



It is important to mention that the argument *cat* is only used for the grah biplot and not in the multidimensional scaling. That is to say, any prior information is giving to perform the analysis of the potential groups that there are in the data, as for example in the logistic regressions, Discriminant Analysis, carts, etc., Therefore, in the case of obtaining a low value of stress and a chart that clearly shows a separation of groups, such as in this example, this would indicate that there is an underlying structure -some clearly differentiated groups- based on the set of measures used.

If access the function metaMDS with *argument METAMDS*, it is possible to change the index of dissimilarity which is used in the Multidimensional Scaling, with the argument *distance*. Experiment with different types of index of dissimilarity is advisable, as it can get a value of less stress and a graph where the groups are more clearly separated. In case of obtaining this error *some dissimilarities are negative*, this may be due to the fact that there are rows or columns in which all values are zero.

Value

A biplot graph is obtained with the dimensions I and II of the multidimensional scaling.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: `http://CRAN.R-project.org/package=car`.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: `http://CRAN.R-project.org/package=IDPmisc`.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Henry, M., Stevens, H. & Wagner, H. (2013) Community Ecology Package. R package version 2.0-10. Available at: `http://CRAN.R-project.org/package=vegan`.

Examples

```
## Not run:
data(ZXIII3)
XIII3(data=ZXIII3, var=c("M2", "M3", "M4", "M5", "M6", "M7",
"M10", "M11",
"M12", "M13", "M14", "M15", "M16", "M17", "M19", "M20", "M21",
"M23", "M24",
"M25", "M26", "M27", "M30"), cat="Family", ellipse=TRUE)

## End(Not run)
```

XIII4 *CANONICAL CORRELATION*

Description

A canonical correlation is applied.

Usage

```
XIII4(data, varX,  varY, labels=NULL, labelsvar="varX", log=FALSE, dim=c(1,2),
expand=TRUE, ellipse=FALSE, convex=FALSE, ResetPAR=TRUE, PAR=NULL, CANCOR=NULL,
SCATTERPLOT=NULL, COLOR=NULL, PCH=NULL, XLIM=NULL, YLIM=NULL, XLAB=NULL, YLAB=NULL,
TEXTX=NULL, TEXTY=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, arrows=TRUE, larrow=0.95,
colAX="black", colAY="blue", file1="Output.txt", file2="xcoef.csv", file3="ycoef.csv",
file4="xscores.csv", file5="yscores.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

`data` Data file.

`varX` Independent variables.

`varY` Dependent variables.

`labels` Variable that allow to tag cases.

`labelsvar` The cases are represented using scores of the independent (*X*) or dependent (*Y*) variables.

`log` If TRUE, the logarithm is applied to the dependent and independent variables.

`dim` Dimensions that are represented on the graph.

`expand` If TRUE, the coefficients of the variables *X* and *Y* and scores of the cases of the independent (*X*) or dependent (*Y*) variables are adjusted to the same scale, when plotted on the graph. This allows a better view of the adjustment of the coefficients and scores on the canonical space, i.e., to see which variables and scores coincide in space generated by the canonical variables, which were created as a linear combination of dependent and independent variables.

ellipse If it is TRUE the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *labels* are depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*.

convex If it is TRUE the convex hull is depicted for each category.

ResetPAR If it is FALSE, the default condition of the function PAR are not placed and those defined by the user on previous graphics are maintained.

PAR It accesses the function PAR which allows to modify many different aspects of the graph.

CANCOR It accesses the function cancor of the candisc package.

SCATTERPLOT It accesses the function scatterplot of the car package.

COLOR It allows to modify the colors of the graph, but they must be as many as different groups the variable *labels* has.

PCH Vector with the symbols on the chart, which must be as many as different groups the variable *labels* has. If it is NULL, they are automatically calculated starting with the symbol 15.

XLIM, YLIM Vectors with the limits of the axes *X* and *Y*.

XLAB, YLAB Legends of the axes *X* and *Y*.

TEXTX It accesses the function to change the labels of cases in graph *biplot*.

TEXTY It accesses the function to change the labels of the coefficients of the dependent variables (*Y*) in the graph *biplot*.

LEGEND It allows to include a legend.

AXIS It allows to add axes.

MTEXT It allows to add texts in the margins of the graph.

arrows If it is TRUE the arrows are shown in the scatterplot.

larrow It modifies the length of the arrows.

colAX, colAY Colors of emphX and *Y* arrows.

file1 TXT FILE. Name of output file with test results.

file2 CSV FILE. Output file name with the coefficients of the variables *X*, shown in the graph, but without the correction applied in the case of the argument *expand=TRUE*.

file3 CSV FILE. Output file name with the coefficients of the variables *Y*, shown in the graph, but without the correction applied in the case of the argument *expand=TRUE*.

file4 CSV FILE. Name of the output file with the coordinates of the cases considering the variables *X*, shown in the graph, but without the correction applied in the case of the argument *expand=TRUE*.

file5 CSV FILEs. Name of the output file with the coordinates of the cases considering the variables *Y*, shown in the graph, but without the correction applied in the case of the argument *expand=TRUE*.

`na`	CSV FILES. Text used in the cells without data.
`dec`	CSV FILES. It defines if a comma"," o a dot "." can be used as decimal separator.
`row.names`	CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

XIII. FACTORIAL MULTIVARIATE ANALYSIS

XIII.4. CANONICAL CORRELATION

The canonical correlation analysis is a generalization of the regressions. A simple regression correlates an independent variable with another dependent.

A multiple regression does the same thing with more than one independent variable and a dependent. The canonical correlation maps a group of variables $x_1, ..., x_m$ with other group $y_1, ..., y_n$. The method is to find two variables $A = a_1x_1 + ... + a_mx_m$, $B = b_1y_1 + ... + b_ny_n$ among which the correlation is maximum. These variables are called composite canonical variables. The analysis calculates several canonical correlations, each with a pair of canonical variables *A* and *B*. In particular, this will calculate many canonical variables as the number of variables the smallest group has (the *x* or *y*), although only the first has practical interest. Often a few pairs of canonical variables make possible to analyze the relationship between the two groups, facilitating the study of our data by reducing the dimension.

This type of multivariate analysis shares aspects with the Principal Component Analysis and factor analysis. But unlike these, seeking internal relationships among the variables of a group, the canonical correlation is looking for a relationship between two sets of variables.

The general requirements of this type of analysis are the same as for a factorial analysis:

1. Qualitative variables are excluded.

2. Relationships between variables must be linear.

However, it is important to point out some characteristics that, although they are not prior hypotheses, it is desirable that these be fulfilled particularly in this analysis:

1. The method works better when the correlation between variables within each group is small, and is large between variables of different group.

2. The variables within a group, although different, must be homogeneous with respect to the type of information, or the topic, to which they relate (remember that in the factorial analysis there is also this requirement, but there are no two groups); otherwise, if the information they contain is not related, the construction of new variables (canonical) as a linear combination of old variables loses meaning.

In short, this type of analysis will be particularly useful when there are two groups of variables, each of which would provide information about a topic, as measured by different variables. Then, it intends to find a correlation between these two groups, or items of information. This method may be more appropriate to compare the two variables to two, and in addition, the correlation found between the canonical variables calculated is always higher than the existing one in any of these two to two comparisons.

FUNCTIONS

The Canonical Correlation was performed with the cancor function of the candisc package (Friendly, 2007; Firendly and Fox, 2013). To perform the *biplot* graph the scatterplot function of the car package was used (Fox et al., 2014). The arrows are depicted with the function Arrows of the package

IDPmisc (Locher & Ruckstuhl, 2014). The convex hull is estimated with the function chull of the package grDevices.

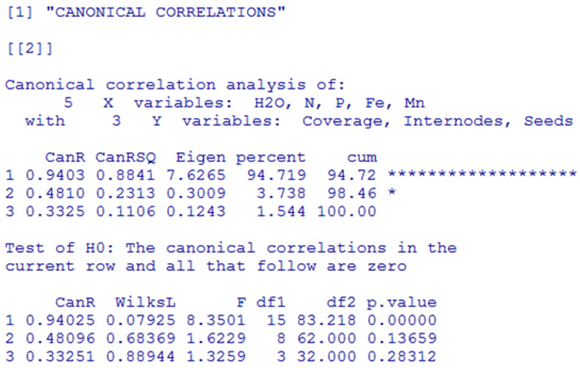
EXAMPLE

This analysis will be used to check if a number of variables, which generally relate to the quality of a soil for the growth of plants, predict the best or worst condition in which the plants that have grown on that soil can be. There is a group of variables that provide information on soil quality. These are as follows: water content (in % of the weight of the soil), nitrogen, phosphate, and a number of trace minerals (in micro moles per gram of soil). The other group of variables provides an idea of condition of the plants, and these are a series of independent parameters related to the growth and reproduction: leaf coverage (in m^2), the average length of internodes (in cm) and the % of the total viable seeds produced. In this case there are data from a set of trees of the same age which have grown in different soils. We want to know if both groups of variables are correlated.

It is very important to keep in mind the advisability of including variables that have low variability, i.e., that all values are equal except a few cases. These variables with little variability can have a great weight in the canonical axes and it should be assessed whether its inclusion is necessary. For example, in the case of working with an abundance of species, it may be that those which are not present in the majority of the samples and appear once or twice have an important weight in the canonical axes and it is necessary to assess whether it is proper to include them.

In the script, the group of independent variables with the argument *varX* and dependent with *varY* are selected. It is s specified by the argument *labels= "Area"* that the scores of cases display differenciating by the sampling areas.

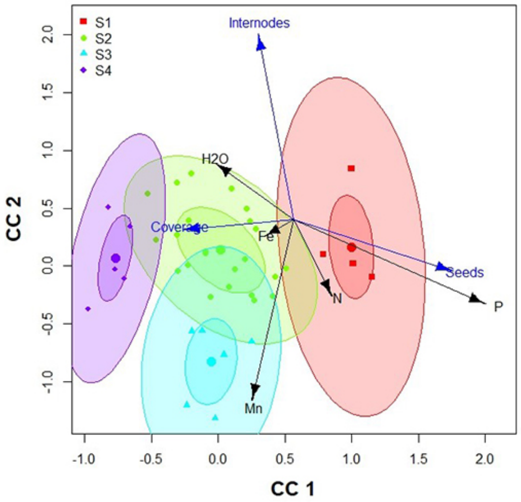
The following table shows the results of the analysis. Three pairs of canonical variables are calculated, because the group of the dependent variables has the fewest number of variables, which are three. The first pair of canonical variables explain the greater portion of the variance (94.72%), and the remaining pairs explain, successively, as much as possible of the remaining variance. To explain different aspects of the variance, the pairs of canonical variables are never correlated with each other.The first canon R (*Canr 1= 0.94*) is equivalent to R's global analysis. The second R canon (*Canr 2 = 0.48*) is independent of the first, the third R canon (*Canr 3 = 0.33*) is independent of the first and the second, and so on in the event that there are more canonical pairs. It is noted that the low value of R, and in fact, by deleting the first pair of canonical variables calculated, R is no longer significant (p = 0.136). Therefore, the correlation is only meaningful in the first calculated couple of canonical variables, that is to say, for the first canonical shaft (p < 0.001). Overall it may be said that the quality of the soil is significantly connected with the condition of the plants with the variables used.



```
[1] "CANONICAL CORRELATIONS"

[[2]]

Canonical correlation analysis of:
      5   X  variables:  H2O, N, P, Fe, Mn
  with    3   Y  variables:  Coverage, Internodes, Seeds

    CanR CanRSQ  Eigen percent    cum
1 0.9403 0.8841 7.6265  94.719  94.72 *******************
2 0.4810 0.2313 0.3009   3.738  98.46 *
3 0.3325 0.1106 0.1243   1.544 100.00

Test of H0: The canonical correlations in the
current row and all that follow are zero

     CanR  WilksL      F df1    df2 p.value
1 0.94025 0.07925 8.3501  15 83.218 0.00000
2 0.48096 0.68369 1.6229   8 62.000 0.13659
3 0.33251 0.88944 1.3259   3 32.000 0.28312
```

In Figure III. 9 the coefficients of the dependent and independent variables are shown, as well as scores of differentiated cases in function of the sampling area. The higher the absolute value of the coefficient, the greater the contribution of that variable to the canonical variable. In this example is of special interest to see what are the variables that have more weight in the first canonical correlation (*CC1*), because as it was seen, this is the only significant one. It is noted that the independent variables of more weight are the phosphate and nitrogen. In what refers to the dependent variables, it will also be noted, for the first canonical correlation, that the variable of more weight is the percentage of viable seeds, which is greater in soils with more concentration of phosphate and nitrogen.

Figure XIII.9. I and II canonical axes of the Canonical Correlation applied to the data of growth and reproduction of plants and variablesthat measure soil quality.



The ellipses show the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each of the sampling sites. It is an indicator of the degree of overlap between the sampling sites.

These levels can be modified by entering in the scatterplot function using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*. With some examples it is observed that it is not possible to estimate the ellipses and this error *Error in svd(sqrt(w/sum(w)) * X, nu = 0):* is obtained or this other *Error in cov.trob(cbind(x[use], y[use]), wt = weights[use]): no positive weights*. In case of obtaining these errors or simply not wanting to show the ellipses, again using the argument emphellipse=FALSE.

The figure shows that the sample areas can be differentiated in terms of the characteristics of the soil and the condition of the plants. Site 1 is where there are higher concentrations of phosphate and nitrogen in the soil and, as a result, a greater percentage of viable seeds. Site 4, on the contrary, it is where the percentage of viable seed is smaller, in principle caused by a soil of a lesser quality, that is to say, the lowest concentrations of phosphate and nitrogen.

It is possible to obtain a greater percentage of explanation in the first axes, and they are even more significant, if those variables that contribute little to the canonical axes are not included in the analysis that is to say, they are in the center of the graph. Therefore, it is advisable to carry out various tests by eliminating variables that have little weight in the analysis and see if this increases the percentage of explanation and/or degree of significance of the canonical axes.

Value

A graph *biplot* with the Canonical Correlation, a TXT file with the test results and four CSV files are obtained: the coefficients of the variables *X* and *Y* and the coordinates of the cases considering the variables *X* and *Y*.

References

Friendly, M. (2007). HE plots for Multivariate General Linear Models. *Journal of Computational and Graphical Statistics*, 16: 421-444.

Friendly, M. & Fox, J. (2013) Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.6-5. Available at: http://CRAN.R-project.org/package=candisc.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: http://CRAN.R-project.org/package=car.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: http://CRAN.R-project.org/package=IDPmisc.

Examples

```
## Not run:

data(ZXIII4)

XIII4(data=ZXIII4, varX=c("H2O","N","P","Fe","Mn"), varY=c("Coverage",
"Internodes","Seeds"), labels="Area", ellipse=TRUE)


## End(Not run)
```

XIV1 *CLASSIFICATION AND REGRESSION TREES (CARTs)*

Description

A Classification and Regression Tree is performed (CART).

Usage

```
XIV1(data, cat, var, ResetPAR=TRUE, PAR=NULL, CART=NULL, RPART=NULL,
file="Output.txt")
```

Arguments

`data`	Data file.
`cat`	Dependent variable.
`var`	Independent variables.
`ResetPAR`	If it is FALSE, the default condition of the function PAR are not placed and those defined by the user on previous graphics are maintained.
`PAR`	It accesses the function PAR that allows to modify many different aspects of the graph.
`CART`	It accesses the function rpart.
`RPART`	It accesses the function rpart.plot.
`file`	TXT FILE. Output file name with the results.

Details

XIV. MULTIVARIATE ANALYSIS OF CLASSIFICATION

It is the set of multivariate analysis methods whose main objective is to group or classify the elements of a sample of objects described by several variables. Compared to the factorial methods, classification techniques generally have an algorithmic approach (not algebraic), and most of them are not based on the linear model. Historically these arise with the computer and informatics, and benefit from the extraordinary power of calculation that these tools provide. There are hundreds of methods of classification or cluster analysis, of which only the most important will be seen.

In general it is to get homogeneous groups (classes, clusters), i.e., constructed in such a way that the elements in each group are similar among them, the classes must be separated or different, so that the elements of different groups are quite different. It is to find a structure in the data that will help to better understand the reality of the studied population.

Many classification methods allow to use variables of various kinds, quantitative or qualitative, nominal or ordinal, and in general other restrictive assumptions such as normality and homogeneity of variances that are usual in other statistical techniques, are not necessary. A prior knowledge of the data is not necessary.

For these reasons the classification methods are often used for exploratory purposes, to establish a first approach to the data, and to generate ideas, raise hypotheses that can be contrasted with other techniques, or to initiate the construction of explanatory models.

XIV.1 CLASSIFICATION AND REGRESSION TREES (CARTs)

The classification and regression trees are a non-parametric method of predicting a dependent variable based on a set of independent variables (Breiman et al., 1984). Its non-parametric nature does not require any hypothesis relating to the distribution of the dependent and independent variables, nor to the relationship between them and their possible interactions. The dependent variable can be categorical (Classification Trees), different species such as, rejection of a treatment, presence/absence of cancer, etc.

The dependent variable can also be continuous (Regression Trees), as for example the body mass index, blood glucose, etc. , in both cases, trees of Classification and Regression (CARTs), the aim is to identify the variables that best identify the dependent variable. For example, to determine which are the clinical, genetic and/or environmental features that can be used to predict if a person can have a certain cancer, or be prone to a heart attack, etc.

Amigo et al. (2007) applied CARTs to the body mass index (BMI) in children and determined that the mother, the consumption of lipids and the hours of television on Sunday, were the variables that best classified children depending on their BMI. A CARTs analysis generally consists of three steps (Guisande & Vaamonde , 2012):

1. **Tree Construction**. In the tree, the root node represents the entire population. This root node is divided into two subgroups on the basis of the partition of a independent or predictor variable. Children nodes are divided by means of the partition of a variable, that can be the same as before or a new one. The process is successively repeated until a stopping condition is met. The divisions are selected so that the "impurity" or heterogeneity of the children nodes would be lesser than that of the parent node. The function of impurity or partition criterion is a measure for determining the quality of a child node, and decides how the partition of a node is carried out in their two children nodes. There are different types of indices to quantify the impurity of the partitions (Breiman et al., 1984).

2. **Tree Pruning**. The tree that has been built is generally over-fitted, i.e., it contains a large amount of levels and a greater complexity does not necessarily mean a better classification. Let us imagine that we are working with different species and want to determine which are the variables that allow distinguish them better. The tree can continue dividing into nodes until the last individual is classified, using a larger number of variables at each step and a more complex model to achieve a zero percentage of poorly sorted individuals. With the increasing complexity of a model, a better setting is always obtained, but sometimes simply very specific peculiarities of the data are collected, which are not useful for describing the behavior of the new data (data validation). Pruning consists of, once built the tree, deleting backwards partitions that do not represent a significant increase in the total capacity of prediction.

3. **Tree validation**. As pruning the complexity of the tree is reduced but the number of individuals poorly classified can increase. The best tree will be one that has the optimum ratio of the rate of poor classification (ratio between the observations wrongly classified and the total number of observations) and the complexity of the tree.

The selection of the best tree for validation is performed. The procedure consists in separating a portion of the sample, which is not used for the construction of the tree, but only for the prediction, measuring the percentage of success in it. It is also often use a cross-validation method, which is to divide the sample into several mutually exclusive groups (usually 10 groups) of approximately equal size, and make the tree leaving a group outside, which is used for the prediction and measurement of the degree of success. The process is repeated until all the groups are used and finally the tree with the least aggregate rate of misclassification (average of the ten groups) is selected.

FUNCTIONS

For the estimation of the CART, the function rpart of the package rpart (Therneau et al., 2014) is used and for its representation, the function rpart.plot of the package rpart.plot (Milborrow, 2014) is used.

EXAMPLE

The example uses demographic data of 2010 of the 19 regions or Autonomous communities in Spain published by the National Institute of Statistics (`http://www.ine.es`). The goal is to determine if it is possible to classify the different regions based on the following demographic variables: number of children per mother, average age of the mother when she has her first child, average percentage of unmarried women with children, the life expectancy of men and women, and the number of births, population mortality and infant mortality (children under 1 year), the last three demographic parameters per 1000 inhabitants.

The dependent variable is the regions and the independent variables or predictors are all the demographic indicators. With the argument *method*, the method of partitioning is defined; if the dependent variable is categorical, as in our example, the *method='class'* will be placed. The method can be omitted, that is what has been done in the example, in which case the function itself decides the most appropriate method depending on the type of dependent variable.

With the argument *control*, and using the function *rpart.control()*, important aspects of the tree can be modified. With the argument *minsplit* the minimum number of observations that must exist in a node are specified for the partition, i.e., as smaller the tree, this will have more nodes. With *minbucket* the minimum number of observations in a terminal node is defined and, therefore, like the previous argument, as it is smaller, it goes more deeply into the tree, i.e., has more nodes and is more complex.

With *cp* the complexity is controlled and allows to save time in the process of pruning, which is the fraction (default 0.01) in which the adjustment indicator must improve to continue the construction of the tree. If the improvement is less than *cp* the process stops. If the user decides to delve deeper and make the tree very complex (*cp* small), or have a more simple tree (*cp* larger).

With *surrogatestyle* the criterion for choosing the best substitute variable for the partition on each node is controlled. If it is 0 the variable with the largest number of items rated correctamentees used, and if it is 1 that has the highest percentage of elements correctly classified on the number of valid cases of that variable; the first option penalises the variables that have missing values.

In the function rpart.plot with the argument *type* the format of the chart is selected, i.e., the way the information comes from the selected variables in each partition and the predominant categories in each node. With the argument *extra* the additional information in the nodes can be displayed, as in this example, it is requested that the number of cases correctly classified as opposed to total are shown on each node.

With *varlen=0* is defined that when the texts of the nodes are complete they are not cut. With *ycompress=TRUE* is defined that when labels overlap, they must move vertically to avoid the overlap.

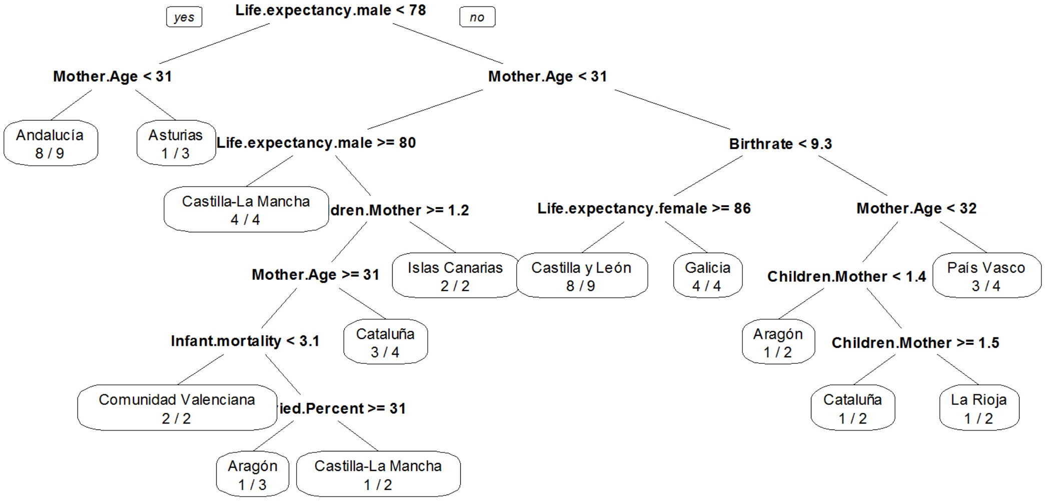
For more details see the arguments on the help menus of functions rpart and rpart.plot and/or Guisande & Vaamonde (2012).

CART is shown in Figure XIV.1. The numbers indicate the provinces that have been correctly classified and the total number of provinces in each region. For example, there are 8 provinces of the region of a total of 9 in the group that shows Castilla and Leon. There is a group that displays 8 provinces of Andalusia from a total of 9.

In general it is observed that many regions are well identified. For example, in the Canary Islands, Galicia, Castile La Mancha and Valencia, all the provinces of the region are within the same group. In others, such as Andalusia, Catalonia, Basque Country, etc., almost all provinces are in the same group. This means that there are demographic differences between regions in Spain, as it is possible to identify relatively well each of the Regions on the basis of these demographic indicators.

The first variable is the life expectancy of men, which separates a group of regions with men who have a life expectancy of less than 78 years to the left and more than 78 to the right. The average number of children per mother and mother's age when she has the first child are also important variables as they will appear in many nodes. The number of births per 1000 deaths is also an important variable indicating that regions grow at different rates. In the tree, it has been chosen not to identify all cases properly, that is to say, there are terminal groups where there are cases of different categories, to make the tree not very complex. What is important in this example is to see if the demographic indicators differ to the regions, and which of these indicators are the most important for this.

Figure XIV1.1. Demographic parameters that best differenciate the regions of Spain.



Value

A graph with the Classification and Regression Tree and a TXT file with the results of the CART are obtained.

References

Amigo, H., Bustos, P., Erazo, M., Cumsille, P. & Silva, C. (2007) Factores determinantes del exceso de peso en escolares. Un estudio multinivel. *Revista Médica de Chile*, 57: 353-357.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) *Classification and Regression Trees*. Wadsworth.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Milborrow, S. (2014) Plot rpart models. An enhanced version of plot.rpart. R package version 1.4-4. Available at: `http://CRAN.R-project.org/package=rpart.plot`.

Therneau, T., Atkinson, B. & Ripley B (2014) Recursive Partitioning and Regression Trees. R package version 4.1-8. Available at: `http://CRAN.R-project.org/package=rpart`.

Examples

```
## Not run:

data(ZXIV1)

XIV1(data=ZXIV1, cat="Region", var= c("Children.Mother", "Mother.Age",
"Unmarried.Percent", "Life.expectancy.male", "Life.expectancy.female",
"Birthrate","Mortality", "Infant.mortality"))
```

```
## End(Not run)
```

XIV2 *DISCRIMINANT ANALYSIS*

Description

Two types of discriminant analysis are applied: Linear and Quadratic.

Usage

```
XIV2(data, cat, var, ellipse=TRUE, convex=FALSE, quadratic=FALSE, expand=TRUE,
dimS=c(1,2), ResetPAR=TRUE, PAR=NULL, CANDISC1=NULL, CANDISC2=NULL, CANPLOT=NULL,
SCATTERPLOT=NULL, COLOR=NULL, PCH=NULL, TEXT=NULL, LEGEND=NULL, AXIS=NULL,
MTEXT=NULL, arrows=TRUE, larrow=0.95, colArrows="black", file1="Var loadings-Linear.csv",
file2="Cat loadings-Linear.csv",file3="Table cross-validation-Linear.csv",
file4="Cases cross-validation-Linear.csv",
file5="Table cross-validation-Quadratic.csv",
file6="Cases cross-validation-Quadratic.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

`data` Data file.

`cat` Categories of cases.

`var` Variables included in the analysis.

`ellipse` If it is TRUE the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *cat* are depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*.

`convex` If it is TRUE the convex hull is depicted for each category.

`quadratic` If TRUE, a Quadratic Discriminant Analysis is performed, in addition to the Linear Discriminant Analysis.

`expand` If TRUE, the coordinates of the categories, in the graph *biplot*, are adjusted to the scale of the coordinates of the cases.

`dimS` Dimensions that are represented in the graph *biplot*.

`ResetPAR` If it is FALSE, the default condition of the function PAR are not placed and those defined by the user on previous graphics are maintained.

`PAR` It accesses the function PAR that allows to modify many different aspects of the graph.

`CANDISC1` It accesses the function candisc of the candisc package, performing the one-Dimensional Linear Discriminant Analysis.

CANDISC2 It accesses the function candisc of the candisc package which performs the Two-Dimensional Linear Discriminant Analysis.

CANPLOT It accesses the function plot.cancor of the candisc package, performing the one-dimensional discriminant graph.

SCATTERPLOT It accesses the function scatterplot of the car package, with the graph *biplot* that performs the Two-Dimensional Linear Discriminant Analysis.

COLOR It allows to modify the colors of the graph *biplot*, that should be as many as different groups the variable *cat* has.

PCH Vector with the symbols of the graph *biplot*, that should be as many as different groups the variable *cat* has. If NULL, they are automatically calculated starting with the symbol 15.

TEXT It accesses the function that allows to modify the labels of the cases in the graph *biplot*.

LEGEND It allows to include a legend.

AXIS It allows to add axes.

MTEXT It allows to add texts in the margins of the graph.

arrows If it is TRUE the arrows are shown in the scatterplot.

larrow It modifies the length of the arrows.

colArrows Color de las flechas.

file1 CSV FILE. Name of the output file with the coordinates of the variables in the graph of the Linear Discriminant Analysis.

file2 CSV FILE. Name of the output file with the coordinates of the categories in the graph of the Linear Discriminant Analysis.

file3 CSV FILE. Name of the output file with the table of predictions using the cross validation of the Linear Discriminant Analysis.

file4 CSV FILE. Name of the output file with the group to which each case belongs and the prediction of the Discriminant Analysis using cross validation of the Linear Discriminant Analysis.

file5 CSV FILE. Name of the output file with the table of predictions using the cross validation of the Quadratic Discriminant Analysis.

file6 CSV FILE. Name of the output file with the group to which each case belongs and the prediction of the discriminant analysis using cross validation of the Quadratic Discriminant Analysis.

na CSV FILES. Text that is used in the cells without data.

dec CSV FILES. It defines if a comma "," or a dot "." is used as decimal separator.

row.names CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

XIV. MULTIVARIATE ANALYSIS OF CLASSIFICATION

XIV.2. DISCRIMINANT ANALYSIS

A multivariate statistical technique widely used to characterize classes of elements when there is a certain number of variables is the discriminant analysis, that allows -from an initial classification- to get a new classification based on the observed values of a set of variables. The method consists in the determination of some functions of the original variables, called discriminant functions, which allow to decide in which class each element must be, using as a criterion for assigning the proximity (or similarity) of each element to the different classes or existing groups. It can be used, for example, to identify the geographical origin of a sample of unknown honey, when there are multiple samples of each area, using the contents or concentrations of different types of sugars that have been analytically determined in all the samples.

The construction of the discriminant functions, linear combinations of the original variables, is carried out by finding those that do the maximum separation between groups, which is equivalent to minimize the variability within the classes. The functions obtained enable to discriminate in the best possible way between the elements belonging to the different classes or groups. The allocation of the elements to classes is performed using the Bayes approach: each element is assigned to the class for which it is greater the conditioned probability of belonging by values that take the discriminant functions.

XIV.2.1. Hypothesis

The Linear Discriminant Analysis requires, so that it can be used correctly, the hypothesis of homogeneity of variances, i.e., that the covariance matrices of each class are equal, or that the dispersion is similar in each class for all the variables. The M-box test is applied to test this hypothesis of homogeneity. It is also necessary that the discriminant functions have Normal distribution. Both of these assumptions of normality of residuals and homogeneity of variances were used in the calculation of the probabilities of assignment of the elements for the various classes. However, these do not usually be checked in practice as the tests M-box and contrasts of normality are very sensitive to small deviations and with actual data they are not virtually met. If these are checked, the Linear Discriminant Analysis can never be applied.

The Linear Discriminant Analysis is reasonably robust against a partial default on these hypotheses (enough when these are fulfilled in an approximate way), and that failure is not critical if the hit ratio in the classification (especially in the cross-validation) is high. It is necessary that the hit ratio is high not only in the whole of the elements, but also in each of the groups. In practical terms what is checked is, therefore, the percentage of success in the cross-validation, and it is neither necessary nor advisable to check the assumptions, unless there is clear evidence that they are unfulfilled (by the type of data, for example when the explanatory variables are binary), in which case it may not be correct to use the Linear Discriminant Analysis. If the hit ratio is low, this could be due to the failure of the assumptions, but in that case there is no need to be checked, because what is prudent is to give up when the discriminant analysis classifies wrongly.

If the argument *quadratic=TRUE* performs a Quadratic Discriminant Analysis, which does not require the homogeneity of variances (Venables & Ripley, 2002), in addition to the linear method. The comparison of the percentage of correct cases classfied, by cross-validation, obtained with the Quadratic Discriminant Analysis and Linear could be used as an indicator of how robust the straight-line method is, since if two percentages are similar, this would indicate that the Linear Discriminant Analysis is correct. However, the Quadratic Discriminant Analysis cannot be applied when any category has only a few cases, usually less than 5. In addition, it is frequent for this error *rank deficiency in group ...* when applying the quadratic method, in which case, it would have to specify the argument *quadratic=FALSE*, so that it cannot be done.

XIV.2.2. Step by step method

The step-by-step approach can be used for the discriminant analysis, with which the variables with the greatest capacity of discrimination are successively and automatically selected in accordance with a particular statistical criterion (for example the minimum value of the coefficient Wilks' lambda), which are incorporated into a set of variables that are involved in the analysis. In some steps a variable previously entered can be ruled out, but when other new variables are considered, these will no longer provide the ability to set a significant discrimination. However it is preferable select the variables that should not automatically be used, but with technical or scientific criteria - non-statistical - by applying those that have a priori a direct relationship (for example type of cause-effect) with the classification.

XIV.2.3. Influence of the different independent variables

The solution obtained to classify cases whose group is unknown a priori is generally obtained, i.e. for identification purposes or prediction. It is also used to reclassify the elements or reassign the elements to the different classes when this classification is not reliable or is only indicative, to study the effect of the different explanatory variables on the classification. With the latter objective the standard coefficients of the first discriminant functions are used: the higher the coefficient, the greater the effect of this variable.

Also the so-called "coefficients of structure," or correlation coefficients between the discriminant functions and the independent variables (which have been used for its construction) can be used. Generally, the older structure coefficients in absolute value correspond to the highest coefficients standard, although on occasions the high coefficient of structure corresponds to a standard low ratio, which allows to detect redundant variables, or vice versa, a coefficient of structure under with a high standard ratio, which identifies a confounding variable, which interacts with the remaining ones distorting the results. Usually both types of variables should be excluded from the analysis.

Univariate ANOVAs are often used to compare the hypothesis that the means of each independent variable are the same for all groups. These also allow to measure the importance or capacity of discrimination of the different variables. If the means are equal, the variable in general will have little or no capacity for discrimination.

XIV.2.4. Validation

When the number of variables is large in relation to the size of the sample, a problem of overfitting may appear. The discriminant functions allow to classify with a high degree of success the elements of the sample, but if we apply these same functions to the classification of new elements, which are not used for the construction of the discriminant functions, the degree of success substantially decreases.

In such cases the peculiarities of the displays are being used to properly assign a high number of cases, but the functions are not valid -are not representative- for the population from which comes the sample. It is necessary to perform some type of validation to ensure that the procedure will be valid for the population and not just for the sample. One option is to set aside a portion of the sample (for example 10 items), which are not used to build the discriminant functions, but are assigned together with them to the various classes. If the proportion of elements correctly classified with this sample of validation is also high, the discriminant analysis will be adequate.

Another possibility is the so-called cross-classification leaving one out. A discriminant analysis is carried out excluding the first item in the sample, then this element is assigned using the discriminant functions obtained. Later the same thing is done with each of the elements of the sample. In this way all the elements of the sample are allocated, without using the corresponding element in the calculations of the discriminant functions and, therefore, all data have been used as sample of validation. If the percentage of success with this cross-validation is high and similar, or only slightly lower than, the hit ratio of discriminant analysis with the whole sample, the validation is positive, and discriminant analysis is suitable for such data.

FUNCTIONS

The Linear Discriminant Analysis was performed with the functions candisc of the candisc package (Friendly, 2007; Friendly & Fox, 2013) and lda of the MASS package (Venables & Ripley, 2002; Ripley et al., 2014). The Quadratic Discriminant Analysis was performed with the function qda of the MASS package (Venables & Ripley, 2002; Ripley et al., 2014). The graph with one dimension was performed with the function plot.cancor of the candisc package (Friendly, 2007; Friendly & Fox, 2013). To perform the graph *biplot* the function scatterplot was used of the car package (Fox et al., 2014). The arrows are depicted with the function Arrows of the package IDPmisc (Locher & Ruckstuhl, 2014). The convex hull is estimated with the function chull of the package grDevices.

EXAMPLE 1

The data correspond to a study conducted in lakes in Colombia where the abundance of phytoplankton and the concentration of some nutrients are analyzed. The objective is to determine if it is possible to discriminate between regions on the basis of the concentration of nutrients. In the script the argument *quadratic=TRUE* has been defined to perform the Quadratic Discriminant Analysis.

With the straight-line method, 97.67% of the cases are correctly assigned. However, the cross-validation shows that the percentage of cases correctly allocated is lower (94.19%), although still high.

The Quadratic method displays a percentage by cross validation of the 95.3%, which is very similar to that obtained with the linear method. Therefore, the regions are distinct from depending on nutrient concentrations. The lakes region of the Andes is distinguished by its low concentration of nutrients (Figure XIV.2) and the first discriminant axis explains 78.3% of the variability.

The second axis, which absorbs a 21.7% of the variance and differentiate between the lakes of the Amazon Basin and the Caribbean (Figure XIV.3). The ellipses show the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each of the regions. It is an indicator of the degree of overlap between regions. These levels can be modified by entering the function scatterplot using the *SCATTERPLOT* argument and modifying the argument *levels=c(0.5,0.95)*.

Figure XIV.2. Axis I of the discriminant analysis discriminating regions, depending on nutrient concentration.

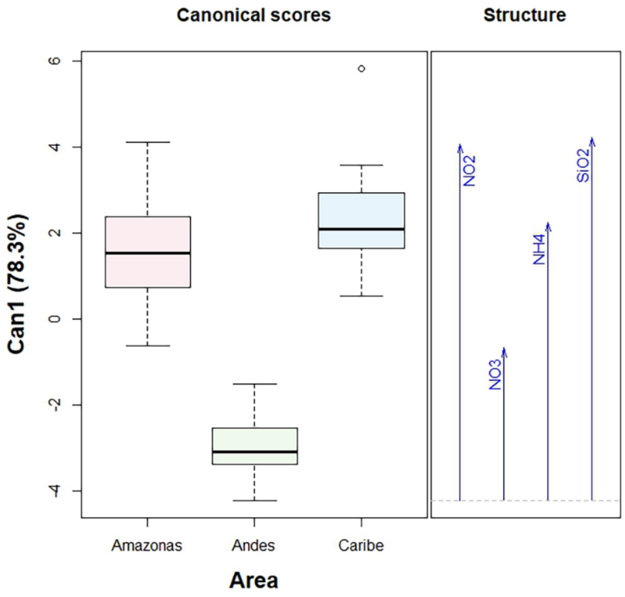
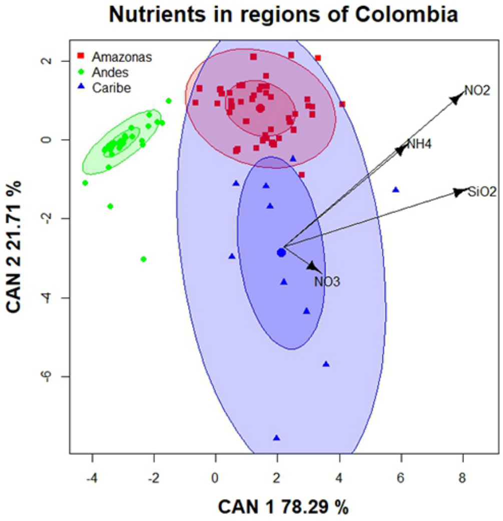


Figure XIV.3. Axes I and II of the discriminant analysis discriminating regions, depending on nutrient concentration.



EXAMPLE 2

In the second example the same previous data are used but the differentiation is carried out at the level of lake instead of region with the argument *cat="Lake"* and the abundance of phytoplankton is also included. In this case the 94.19% is properly identified, and cross-validation is the 90.7% with the linear method. Logically, it is more difficult to differentiate the lakes regions. Although the argument *quadratic=TRUE* has been defined, the function will not automatically applies the quadratic method, as there are categories with few cases. The first axis separates Lake Tota well for its high concentrations of cyanobacteria and Chlorophyceae as well as and low concentrations of nutrients (Figure XIV.4). The second axis, which absorbs a 16.99% of the variance, is mainly determined by the differences in the silicate (S_iO_2) and for the abundance of Chrysophyceae and Dinophyceae (Figure XIV.5).

Figure XIV.4. Axis I of the discriminant analysis applied to species of phytoplankton discriminating lakes depending on the composition of the phytoplankton and nutrients concentration.

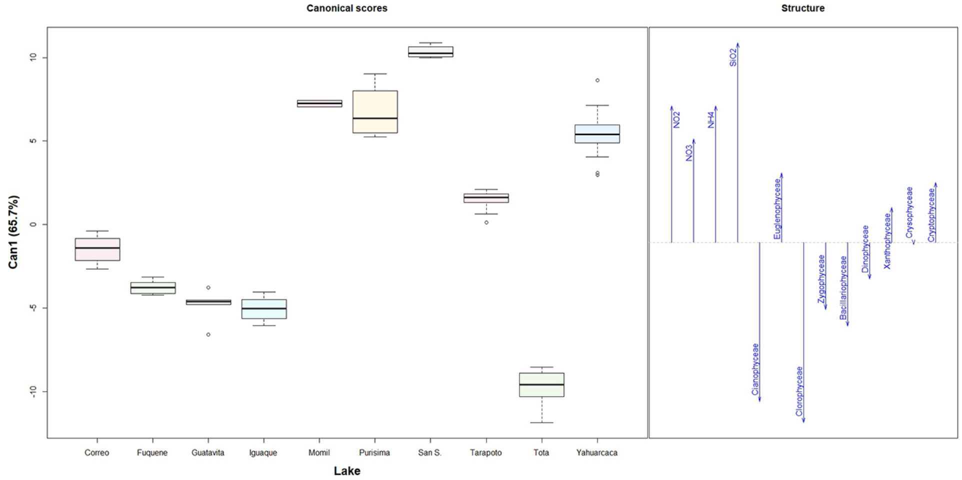
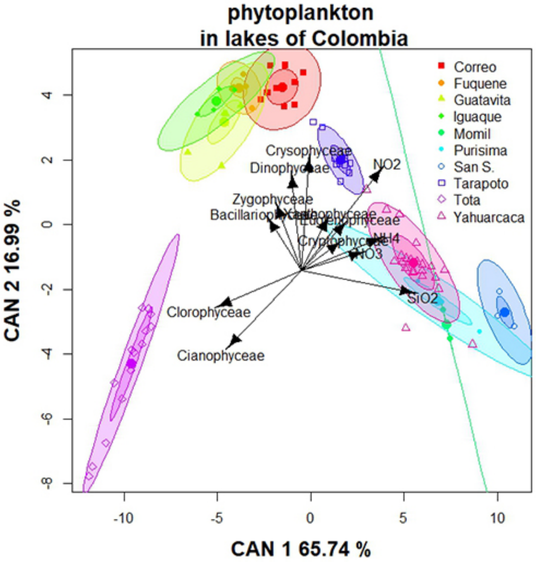


Figure XIV.5. Axes I and II of the discriminant analysis applied to species of phytoplankton discriminating lakes, depending on the composition of the phytoplankton and nutrients concentration.

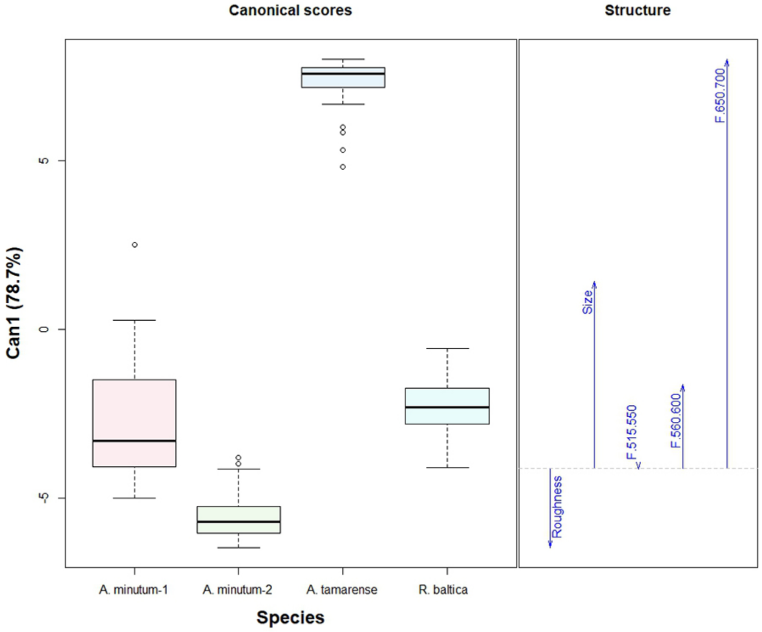


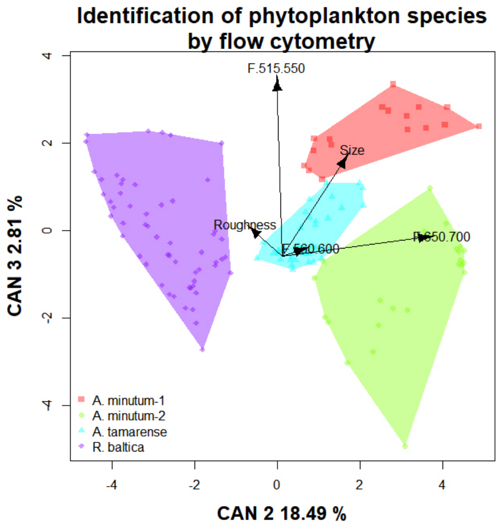
EXAMPLE 3

Data obtained from a flow cytometer of size, roughness and emission of radiation at different wavelengths, from cells that belong to different species of phytoplankton: *Rhodomonas baltica*, *Alexandrium tamarense*, and two strains of *Alexandrium minutum*. The objective is to determine if with the characteristics of the cells obtained in the flow cytometer is possible to discriminate between the different species and strains within the same species. Convex hull instead on the ellipses is estimated with the arguments *ellipse=FALSE* and *convex=TRUE*.

The results show that the 100% of the cases are assigned correctly. The cross validation with the linear and quadratic methods show an identical value, still a high degree of cases correctly assigned (99.29%), in such a way that only a case of *A. minutum-1* is identified as *A. tamarense*. The first axis (Figure XIV.6) explains a 78.7% of the variance and clearly differentiates the species *A. tamarense* for other species, by the emitted radiation between 650 and 700 nm. Sometimes, as shown in the Figure XIV.7 two axes that do not explain a high proportion of the variance, as in this case, the axis II (18.49%) and III (2.81%), can clearly differentiate the groups.

Figures XIV.6. and XIV.7 Axes I and II of the discriminant analysis applied to species of phytoplankton using cell characteristics.





Value

For the Linear Discriminant Analysis a graph with the first dimension is obtained and the *biplot* with the first two dimensions, which can be other option that the user can select in case there are more than two categories. In the window of *Replies* the variance explained by each discriminant function is obtained (only for the Linear Discriminant Analysis) and the percentage of cases correctly identified, also by cross validation, for the two types of discriminant. The CSV files that are obtained for the Linear Discriminant Analysis are: the coordinates of the cases and variables in the axes, a table with a number of classified individuals in each group by cross-validation and a file with the actual group that belongs to each case and the prediction that the model makes. For Quadratic Discriminant Analysis a table with the number of classified individuals in each group by cross-validation is obtained and a file with the real group to which each case belongs and the prediction that the model makes.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: `http://CRAN.R-project.org/package=car`.

Friendly, M. (2007) HE plots for Multivariate General Linear Models. *Journal of Computational and Graphical Statistics*, 16: 421-444.

Friendly, M. & Fox, J. (2013) Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.6-5. Available at: `http://CRAN.R-project.org/package=candisc`.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: `http://CRAN.R-project.org/package=IDPmisc`.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: `http://CRAN.R-project.org/package=MASS`.

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:
data(ZXIV2)

#Example 1. Differences in the physical-chemistry between regions

XIV2(data=ZXIV2, cat="Area", var=c("NO2","NO3","NH4","SiO2"), quadratic=TRUE,
MTEXT=c("text='Nutrients in regions of Colombia'", "side=3",
"line=1", "cex=1.7", "font=2"))

#Example 2. Differences in the abundance of phytoplankton among lakes

XIV2(data=ZXIV2, cat="Lake", var=c("NO2","NO3","NH4","SiO2", "Cianophyceae",
"Euglenophyceae", "Clorophyceae", "Zygophyceae", "Bacillariophyceae",
"Dinophyceae", "Xanthophyceae", "Crysophyceae", "Cryptophyceae"),
quadratic=TRUE, LEGEND=c("x='topright'", "legend=unique(datos3[,1])", "bty='n'",
"col=color1", "pch=pcht"), MTEXT=c("text='Nutrients  and abundance of
phytoplankton\n in lakes of Colombia'", "side=3", "line=0", "cex=1.6", "font=2"))

#Example 3

data(ZXIV3)

XIV2(data=ZXIV3, cat=("Species"), var=c("Roughness","Size","F.515.550","F.560.600","F.650.700"),
quadratic=TRUE, dimS=c(2,3),LEGEND=c("x='bottomleft'", "legend = unique(datos3[,1])",
"bty='n'", "col=color1","pch=pcht"),
MTEXT=c("text= 'Identification of phytoplankton species\nby flow cytometry'",
"side=3", "line=0", "cex=1.7", "font=2"), ellipse=FALSE, convex=TRUE)

## End(Not run)
```

XIV4 *DENDROGRAM*

Description

A hierarchical classification is applied (Dendrogram).

Usage

```
XIV4(data, var, cat=NULL, ResetPAR=TRUE, PAR=NULL, HCLUST=NULL, PLOT1=NULL,
PLOT2=NULL, LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL)
```

Arguments

`data` Data file.

`var` Variables that are included in the analysis.

`cat` Case categories.

`ResetPAR` If it is FALSE, the default condition of the function PAR is not placed and maintained those defined by the user on previous graph.

`PAR` It accesses the function PAR that allows to modify many different aspects of the graph.

`HCLUST` It accesses the function hclust of the stats package.

`PLOT1` It accesses the function that allows to modify the graphical representation of the dendrogram if the argument *cat* is not *NULL*.

`PLOT2` It accesses the function that allows to modify the graphical representation of the dendrogram if the argument is *cat=NULL*.

`LEGEND` It allows to include a legend.

`AXIS` It allows to add axes.

`MTEXT` It allows to add text in the margins of the graph.

`TEXT` It allows to add text in any area of the inner part of the graph.

Details

XIV. MULTIVARIATE ANALYSIS OF CLASSIFICATION

XIV.4. DENDROGRAM

It is one of the most commonly used methods. A sequence of partitions (classifications) hierarchically organized is obtained, from a single class which contains all of the elements, until *n* classes, each of them with a single element. The sequence can be graphically represented by a tree.

Firstly, the algorithm consists of joining the two closest elements among themselves; the resulting class is considered a new element that replaces the previous two. Then the two closest elements that will join in a new class are searched again. The process continues, and the various classes are successively joining in larger classes, until all of the items are eventually a single class.

XIV.4.1. Distance criterion

To decide which are the elements of the sample closest to each other, it is necessary to have a distance criterion. There are different definitions or types of distance and it is necessary to choose the most appropriate for the data used.

Euclidean: is the most used, and corresponds with the usual geometric distance in a multidimensional space. It is defined as the square root of the sum of the squares of the differences in the values of each variable for the two elements. Most of the statistical programs consider the Euclidean distance as an option by default Euclidean distance squared: assigns a greater weight to the more distant objects.

Manhattan Distance: It is the average of the difference between dimensions. The name comes from its relationship with the real distance between two points in a city with streets that are cut at right angles, where in general the straight line is not a possible way.

Mahalanobis Distance: Each dimension is weighted by the inverse of its standard deviation. This distance is an important property: it is invariant to changes in the unit of measure, which makes it especially useful when the variables are heterogeneous.

XIV.4.2. Clustering Agglomeration Criterion

When the process of classification was started, to unite in a single class the two closest elements among themselves, this class passes to be considered a new element that replaces the former two. To continue the algorithm is now necessary to calculate the distance between this new element and all of the above, and for this there are various options, which determine the criterion of agglomeration or conglomeration.

There are different criteria of distance between classes, or between classes and elements, which correspond to the mathematical concept of ultrametric, more restrictive than the classic concept of distance or metric:

Simple link (nearest neighbor): The distance between the two groups is the distance between the two closest objects between the different groups. Sometimes, using this procedure of conglomeration can lead to a "chaining effect of", by which each phase in a class with an isolated element is joined instead of joining two classes between them with several items each, giving rise to a hierarchical sequence of atypical classifications of little practical use.

Complete link (farthest neighbor): The distance between two groups is the longest distance between two objects in both groups.

Mean link or inter-group link: mean distance between all pairs of objects in the two groups.

Intra-group link: average distance between all pairs of objects from the union of two groups. It considers not only the average distances between elements of the two groups, but also the distances between elements of the same group.

McQuitty method: the distance to any object from a group that is a union of two other groups. It is the average of the distances of those two, weighted by the number of units of each group.

Link of centroids: distance between the mean or centroid of the two groups.

Ward method: In each step, the two elements that give rise to a lower loss of information are joined together, it is considered as the sum of squares of distances of each object at the center of its class (minimum variance or inertia).

XIV.4.3. Dendrogram

The main result of this technique is a graph, called dendrogram or Classification tree, which allows to view the entire classification process since its inception with each element in a class until the end with a single class in which all elements of the sample are. In this graph, all partitions or classifications obtained and their hierarchical relationship are observed. For better interpretation cases or classified items can be labeled.

From the dendrogram, it is possible to choose the partition of interest, depending on the distances between classes, which are measured in the lateral scale. The distance between two classes is given by the length of the branch which unites them in the tree; generally in the tree the greater separation between partitions is sought, that allows to view the most natural classification of all objects.

The algorithm in each step requires the calculation of the matrix of distances between all the elements; when the sample contains a large number of objects (several hundred, or more than a thousand) the procedure is excessively slow, in such a way that many minutes may be required or even hours before reaching the final result.

On the other hand, the graph obtained, dendrogram, is not easily interpretable when it is too large. The method of hierarchical classification is not useful for large samples, although in that case this can still be applied to a part of the data (a random sample of the data).

FUNCTIONS

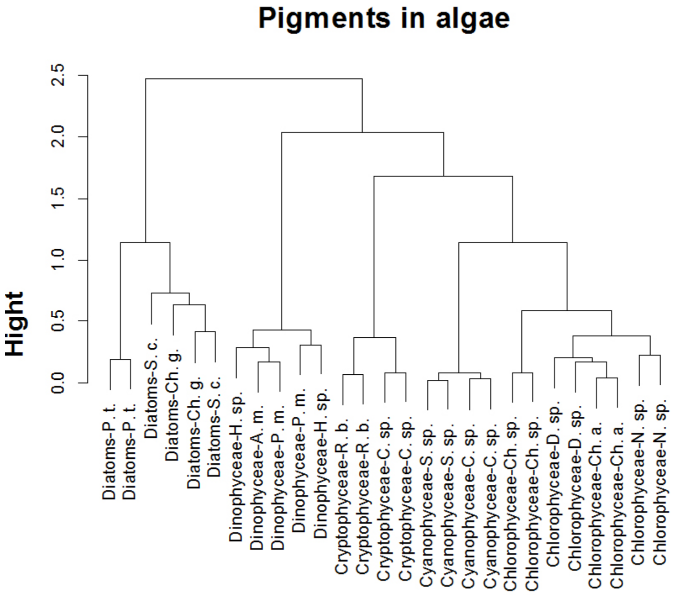
The function hclust of the stats package is used.

EXAMPLE

Data on the relative concentration of 19 pigments, in relation to the concentration of chlorophyll a, phytoplankton species that belong to different classes of algae: Diatoms, Chlorophyceae, Dinophyceae, Cyanobacteria and Cryptophyceae. The objective is to determine if it is possible to differentiate the kinds of algae according to their relative composition of pigments.

In a graph obtained (Figure XIV.9), it is observed that the different classes of algae are perfectly grouped depending on the type of pigments that these have. Diatoms are a group very different from the rest, as well as cryptophyceae and dinophyceae algae.

Figure XIV.9. Hierarchical classification of species of phytoplankton from its pigment composition.



Value

A graph called dendrogram is obtained.

Examples

```
## Not run:

data(ZXIV4)

XIV4(data = ZXIV4 , cat = "Species" , var = c("Chlorophyll.c2", "Chlorophyll.c1",
"Methyl.chlorophyllide.a", "Peridin", "Fucoxanthin", "cis.neoxanthin",
"Violaxanthin", "Diadinoxanthin", "Dinoxanthin", "Alloxanthin", "Zeaxanthin",
"Lutein", "Chlorophyll.b", "Chl.b.epir", "Chlorophyll.a.alllomer",
"Chlorophyll.a.epimer", "beta.u.carotene", "beta.e.carotene", "beta.beta.carotene"),
PLOT1 = c("main='Pigments in algae'", "sub=''", "xlab=''", "ylab='Hight'"))


## End(Not run)
```

XIV5 *HEAT MAP*

Description

A heat map is applied.

Usage

```
XIV5(data, cat, var, ResetPAR=TRUE, PAR=NULL, HEATMAP=NULL, mean=TRUE,
LEGEND=NULL, AXIS=NULL, MTEXT= NULL, TEXT=NULL)
```

Arguments

data	Data file.
cat	Case categories.
var	Variables that are included in the analysis.
ResetPAR	If it is FALSE, the default condition of the function PAR is placed and maintained those defined by the user on previous graph.
PAR	It accesses the function PAR that allows to modify many different aspects of the graph.
HEATMAP	It accesses the function heatmap of the stats package.
mean	If TRUE, the heat map is performed using the average value of each of the groups of the variable *cat*, instead of using all cases.
LEGEND	It allows to include a legend.
AXIS	It allows to add axes.
MTEXT	It allows to add text in the margins of the graph.
TEXT	It allows to add text in any place of the inner part of the graph.

Details

XIV. MULTIVARIATE ANALYSIS OF CLASSIFICATION

XIV.5. HEAT MAP

The heat map shows, with different colors and tones, the intensity of the relationship between the groups defined by the user. This applied to an array of data, with cases and variables (rows and columns), sorts both simultaneously developing a marginal dendrogram for the rows and another for the columns, jointly sorted. The colors indicate the intensity of the relationship or the values of the variable represented, from white (maximum ratio or highest value), passing by the yellow and orange to deep red (minimum value), although these colors by default can be freely modified by the user.

FUNCTIONS

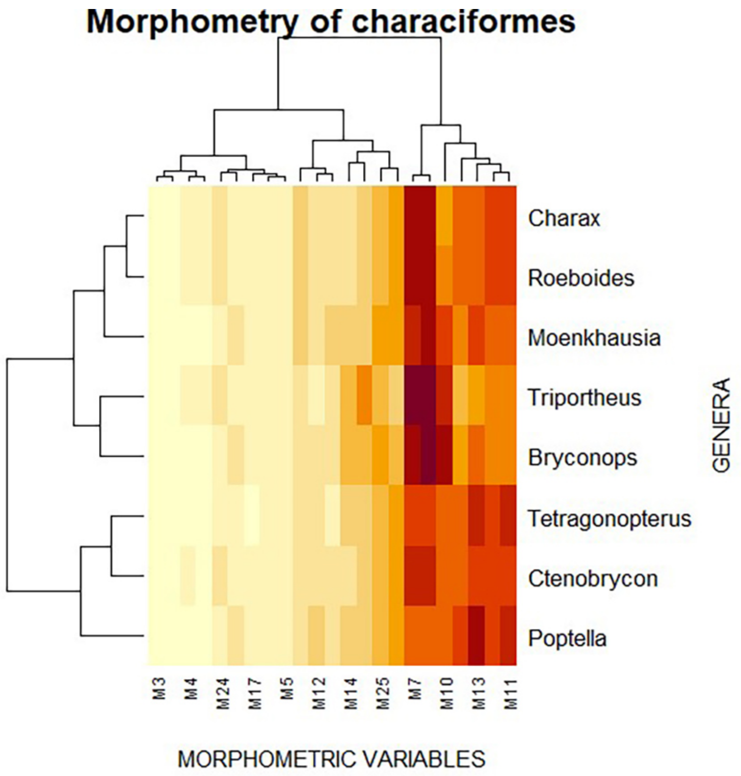
The function heatmap of the stats package is used.

EXAMPLE

Morphometric data from various species of the order Characiformes, such as the length of the base of the dorsal fin (M12), height of the body (M11), and so on. For more details see Guisande et al. (2010). The objective is to determine which are the variables that best differentiate among genera.

In the script, the argument *margin* is a vector with 2 numbers that specify the margins left for the rows and columns.

Figure XIV.10. Heat map of the morphometric variables that best differentiate among genera of the Order Charciformes.



In a heat map variables that have the same hue, that is to say, that are of the same color throughout the gradient, they are the least affecting the classification. On the morphology of the fish, the variables that less influence in the classification are the group of variables on the left, of the M3 to M5, which form a distinct group in the dendrogram of the top (columns). This group of variables contribute less to distinguish the genus of fish, since they all have almost the same shade and, therefore, form a single rectangle, almost in its entirety of red color, on the left side of the graph. The group of variables of the M21 to M19 allow greater discrimination than the old ones, because they have a greater variety of wholes. The variable M7 and M6 also contribute to differentiate the genera, and the fact that are differentiated in a separate block seems to indicate that discriminate well to any particular genus. There are also differences among genera, in the group of variables on the right, of the M10 to the M11.

Value

A heap map graph is obtained.

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

Examples

```
## Not run:

data(ZX3)

XIV5(data=ZX3,  cat="Genus", var=c("M2", "M3", "M4", "M5", "M6", "M7" ,"M10",
"M11", "M12" ,"M13", "M14", "M15", "M16", "M17", "M19", "M20", "M21", "M23",
"M24", "M25", "M26", "M27", "M28"), HEATMAP=c("margins=c(5,10)",
"xlab ='MORPHOMETRIC VARIABLES'", "ylab= 'GENERA'", "main = 'Morphometry of characiformes'"))


## End(Not run)
```

XV1 *ARIMA MODEL*

Description

Manual and automated ARIMA models are applied.

Usage

```
XV1(data, varY, varX=NULL, frequency, start, order=NULL, seasonal=NULL,
ResetPAR=TRUE, PAR=NULL, TS=NULL, PLOTFORECAST1=NULL, PLOTFORECAST2=NULL,
TSDISPLAY=NULL, NDIFFS=NULL, ARIMA=NULL, AUTOARIMA=NULL, FORECAST=NULL,
LINES=NULL, SEASONPLOT=NULL, file="Output.txt")
```

Arguments

`data`	Data file.
`varY`	Dependent variable.
`varX`	Independent variables.
`frequency`	Number of observations per unit of time.
`start`	The date of the first observation. It can be a single number or a vector of two integers, specifying a natural time unit.
`order`	Non-seasonal component of the model. Vector with three numbers (*p*, *d*, *q*) that correspond to the auto regressive component (*p*), integration component (*d*), and moving average component (*q*).
`seasonal`	Seasonal component of the model. Vector with three numbers (*p*, *d*, *q*) that correspond to the auto regressive component (*p*), integration component (*d*), and moving average component (*q*).
`ResetPAR`	If it is FALSE, the default condition of the function PAR are not placed and those defined by the user on previous graphics are maintained.
`PAR`	It accesses the function PAR that allows to modify many different aspects of the graph.
`TS`	It accesses the function ts that allows to modify the parameters of the time series.
`PLOTFORECAST1`	It accesses the function plot.forecast that allows to modify the graphs of the variables over time.
`PLOTFORECAST2`	It accesses the function plot.forecast that allows to modify the forecasting graph.
`TSDISPLAY`	It accesses the function tsdisplay which allows to modify the residual plot and the graphs of the simple autocorrelation function (*ACF*) and partial correlation coefficient (*ACF* partial or *PACF*).
`NDIFFS`	It accesses the function ndiffs that calculates whether the series is stationary or non-stationary.
`ARIMA`	It accesses the function Arima that estimated the manual ARIMA model.
`AUTOARIMA`	It accesses the function auto.arima that estimates the automated ARIMA model.
`FORECAST`	It accesses the function forecast that performs the predictions of the model.
`LINES`	It accesses the function lines that allows to modify lines from graph lines used to make predictions.
`SEASONPLOT`	It accesses the function seasonplot representing the seasonal chart.
`file`	TXT FILE. Name of the output file with the results of the analysis.

Details

XV1. TIME SERIES

A time series is a set of observations from a variable taken over regular intervals of time, such as the number of cars sold by a manufacturer each month during the last ten years. The time series appear in virtually all fields of activity. The interest of its statistical analysis lies in the study of the behavior of the series, which helps explain their variations and, above all, the possibility to predict future values.

The fundamental principle that underlies the analysis of time series is that the current values (and future) of the series depend to a certain extent on the values that has taken in the past, so that we might consider that the series has a certain amount of inertia that prevents -or recently probable - overly abrupt changes in short periods of time. That inertia can be used for prediction, admitting that in the future will behave in a similar way as it did in the past. The success in the analysis and prediction of future values will depend largely on the availability of a wide and reliable sample, and the regularity of the series. Unfortunately, many series are dynamic or changing, and do not behave in accordance with rigid formulas, so that the statistical model used is merely an approximation more or less reasonable to the true operation of the series. In such cases the predictions should be done only in the short term, and with a limited degree of reliability.

In a time series the most frequent studies that often are made are the creation of a model and the analysis of the cycles. For more details on time series see Guisande et al. (2011).

XV1.1. ARIMA MODEL

There are many statistical models applied to the time series, among which the ARIMA model stands out for its power and effectiveness (Box & Jenkins, 1976).

At the time of making a model one must take into account the following analysis: the autocorrelation, if the series grows, decreases or is stationary, seasonality, and, finally, see if there are variables which may be related to the dependent variable.

XV1.1.1. Autocorrelation

To know more about the series it is useful to study the correlation or association between the values of the series in each period and the values in the previous periods, i.e., the relationship between the series and itself (autocorrelation) moved one or more periods.

The simple autocorrelation function (ACF) tends to be typically used, which shows the set of coefficients of autocorrelation, and indicates to what extent the variable depends on itself.

XV1.1.2. Stationary and non-stationary series

The ARIMA model assumes as a precondition, that the series is stationary, i.e., that the behavior remains unchanged in the long term, because the whole approach of this methodology is based on the fact that trends in the future will be similar to that which has been in the past (inertia of the series). In particular, it must be verified that its average value remains stable (no trend increasing or decreasing) and its variability is also stable, does not change (no increases or decreases). If these conditions are not met, it is necessary to make any type of transformation of the variable, usually simple, leading to a stationary series. If there is no appropriate transformation, the ARIMA model should not be applied.

Dickey & Fuller (1979) developed a procedure for testing whether a variable has a unit root or, equivalently, that the variable follows a random walk. There is an extension of the Dickey-Fuller test called the augmented Dickey-Fuller test (ADF), which removes all the structural effects (autocorrelation) in the time series. This test was also used to determine whether the series is stationary.

XV1.1.3. Seasonality

The time series, can be seasonal or non-seasonal. Many time series contain seasonal variations, or behavior patterns that are repeated on a regular basis: The number of tourists is consistently higher in summer than in winter, the volume of purchases in a few department stores is highest during the first few days of each month, the number of cars that pass by a street is greater between the 19h and 21h in the range 22h-24h. Therefore, the seasonal component is defined as those movements of the series that are repeated on a regular basis, being the periodicity lower than a year. Seasonality can be related to different periods of time (months, days, hours, etc.), and in some cases it is extremely regular. To detect patterns of seasonality is very useful to graphically represent the series.

The ARIMA models allow you to include a seasonal component, that can be used for forecasting purposes, and which has the same structure (*p*, *d*, *q*) as the non-seasonal part. These components (*p*, *d*, *q*) are explained later, when considering the construction of the ARIMA model.

In those time series in which there is seasonality, it is necessary to correct this effect for a better interpretation of the model, and is resorted to the seasonally or seasonal correction. To perform this operation is necessary to isolate in the first place the seasonal component, which will enable its subsequent elimination.

XV1.1.4. Covariates

In addition, it is possible to include other explanatory variables or covariates in the model that supposedly have a causal relationship with the dependent variable, or at least are closely related to it.

Thus, in addition to identifying possible variables that influence our time series, if the relationship between these covariates and our dependent variable is significant, it should be expected to improve the model, i.e., to be more predictive.

XV1.1.5. Model building

In the ARIMA models series values are obtained through a combination of three components (*AR*, *I*, *MA*) or also called (*p*, *d*, *q*):

1) Auto Regressive Component (*p*). This shows the part of the value of the variable that depends on the values above, in the history of the series. This is obtained by a linear regression model in which the dependent variable is the component *AR* of the current period, and the explanatory or independent variable the value immediately preceding or a certain number *p* from previous values of the same series. For example, the value AR of the month of June depends on the value of the series in the month of May ($p = 1$), or of the values of March, April and May ($p = 3$). The value of *p*, the order of auto regression, must be specified for each application.

2) Integration component (*d*). A series is stationary if its media, its variability, and its structure of correlations do not change with time. If the series is not stationary, but that has an increasing or decreasing global behavior, there is an integration component that must be determined in the first place. The elimination of this component allows to consider the series as stationary, and subsequently it must be integrated through a reverse process to its elimination. For the calculation of the integration component differences between consecutive values of the series are obtained, by repeating this procedure on the differences obtained the necessary number of times *d* (order of integration) to achieve a stationary series.

3) Component of moving average (*q*). In general the series do not behave in the same way in relation to the model, but there is a residual element, that explains the differences between the observed value in each period and the value obtained by the auto regressive model. This element represents the innovation in the series, the part that does not respond to the inertia or the previous historical behavior, but that is new in each period. Analyzing the series we can determine these residuals or terms of innovation in the time periods already elapsed. The component MA of each period is obtained as an average of a certain number *q* of these residuals in previous periods, and *q* the order of moving averages.

In general, the values of *p*, *d* and *q* are small (zero or one). The higher values of these parameters of the model allow a better fit to the data of the sample, but these can also produce an effect of overfitting in the model by introducing specific aspects of the available sample, not representative of the population from which they were obtained but specific to the sample so that the model can be very suitable for the sample, but inappropriate for the population from which it comes. In general, a simple model, that it may be more suitable for the purposes of prediction, is preferred, especially if the sample is not very large, or if the series is not very stationary in their behavior but slightly changing.

FUNCTIONS

The time series is performed with the function ts of the base package stats. The manual ARIMA model with the function Arima and the automated ARIMA model is estimated with the function auto.arima. The graph, representing the variables in the time, it is performed with the function plot.forecast, to represent the simple autocorrelation function (*ACF*) and the partial correlation coefficient (*partial ACF* or *PACF*) the function tsdisplay is used. To calculate if the series is stationary or non-stationary, the function ndiffs is used. The predictions of the model are estimated using the function forecast and the seasonal chart is performed with the function seasonplot, all of them from the forecast package (Hyndman et al., 2018). The lines on the chart of predictions are carried out with the function lines of the base package graphics. The Box-Ljung test is performed with the function Box.test and the normality of residuals is analyzed with the function shapiro.test, both of the base package stats. The function lillie.test of the package nortest (Gross, 2013) is also used to perform the Kolmogorov-Smirnov Normality test with the correction of Lilliefors to the residuals. The augmented Dickey-Fuller test was performed with the function adf.test of package tseries (Trapletti & Hornik 2017).

EXAMPLE

The monthly condition factor (*FC*) of the anchovy (*Engraulis encrasicolus*) over several years is an indicator of the reproductive status of individuals. The objective is to determine the effect of the average monthly values of temperature (in degree C), chlorophyll (in mgm^{-3}) and stability of the water column (in m^3s^{-3}) on the reproductive status of this species in the Strait of Sicily (Basilone et al., 2006). As a *FC* and the rest of variables can covariate over time, it is necessary to quantify the possible effect of time on *FC*, in other words, to enter the potential effect of time on *FC* in the model, to be able to determine the actual effect of the variables listed above on *FC*. Therefore, it is first always useful to analyze the temporal evolution of the series, in this case of *FC*, without taking into account the effect of the independent variables.

It is important to mention that it is not necessary to have complete series, because it makes an estimate of the missing data. However, it is necessary that this row indicates the date to which it corresponds. For example, if the data, in January 1998, is missing, it is necessary to display the row indicating that it is January 1998, and then leaving blank missing columns.

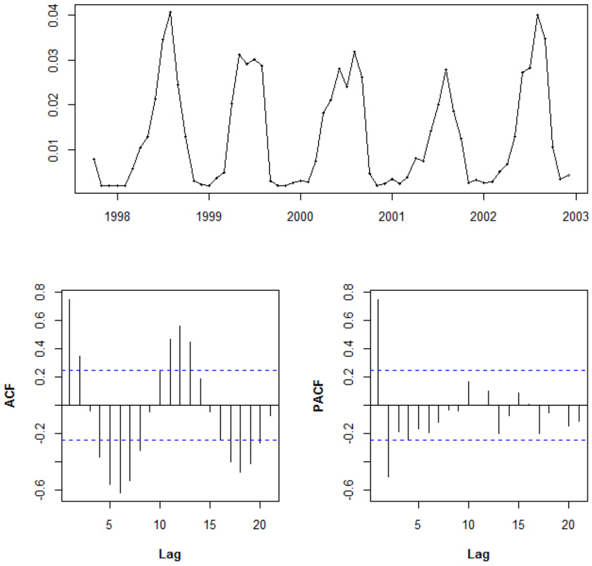
AUTOMATED ARIMA MODEL WITHOUT INDEPENDENT VARIABLES

To be the monthly sampling, in the script in the argument *frequency= 12* the year is considered as a unit of time. The argument *start=c(1997,10)* because the first sampling was in October of 1997. When you do not specify anything in the arguments *order* and *seasonal*, which are the default condition, then the automated ARIMA model, i.e., an ARIMA model is built using an expert modeler.

Autocorrelation

There is a clear annual seasonality (Figure XV1.1), since all years the same pattern is repeated, with low values of *FC* in winter and higher in summer. On the other hand, the series as a whole seems not to show a trend (is horizontal), or cycle.

Figure XV1.1. Temporary representation of the condition factor of the anchovy, and values of ACF and ACFP.



The role of simple autocorrelation (*ACF*), which is shown in Figure XV1.1, shows a set of coefficients of autocorrelation, and indicates the extent to which the variable depends on it. When the value of the series Y_t strongly depends on the previous value $Y_t - 1$ there is a high correlation coefficient A_1 between the series and the same series displaced a period, but given that $Y_t - 1$ in turn, depends on the previous value to the $Y_t - 2$, also the correlation coefficient will be high along with the series shifted two periods, R_2, and so on. The graph shows coefficients that diminish gradually. In this function, it is represented the coefficient corresponding to zero delay, which always worth one (it should not be interpreted, only serves as a reference for comparison).

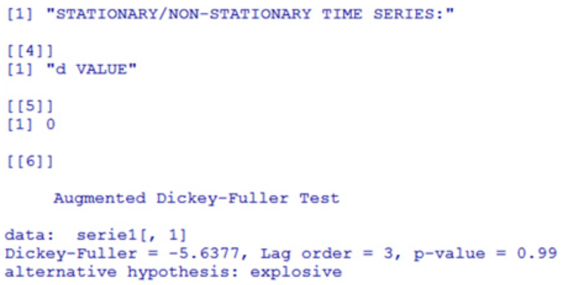
It is possible that a part of such dependence measures the simple correlation coefficient can be indirect, through the intermediate values, so that only the values closest to the current one should be considered (perhaps one or two only). Therefore, the partial correlation coefficient (*partial ACF* or *PACF*) is more appropriate to analyze these relationships of dependence than the simple correlation coefficient, since by eliminating the influence of the remaining variables (intermediate values) this isolates the effect of each delay on the present value of the series, and allows to study the correlation between the series and the same series delayed without including the indirect effects.

In the partial autocorrelation function (Figure XV1.1) we see that only the first two are significant, and the first has a high value. With these results, an appropriate model could have p = 1 or p = 2 and q = 0.

Stationary or non-stationary series

The results show that the series has a value $d = 0$, therefore, the series is stationary. If the series was not stationary then it would be sufficient to put the value of d recommended in the ARIMA model, since what the ARIMA model does is differentiate the series d times, with what the series becomes stationary, and once the model is applied the resulting series integrates d times to undo the change.

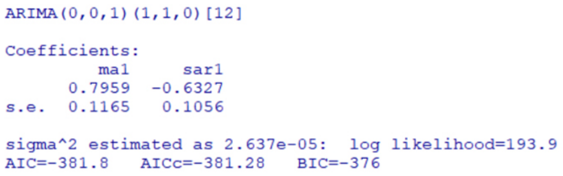
The result of the augmented Dickey-Fuller test (ADF) corroborates that the series is stationary, because the probability is 0.99, so null hypothesis that the time series is stationary is accepted.



```
[1] "STATIONARY/NON-STATIONARY TIME SERIES:"

[[4]]
[1] "d VALUE"

[[5]]
[1] 0

[[6]]

	Augmented Dickey-Fuller Test

data:  seriel[, 1]
Dickey-Fuller = -5.6377, Lag order = 3, p-value = 0.99
alternative hypothesis: explosive
```

Construction of the automated ARIMA model

The function has a procedure that finds the most suitable model between the different values of the parameters *p*, *q*, and *d*, for both the regular part as to the seasonal one. This uses the algorithm of Hyndman & Khandakar (2008), that choose the parameters by minimizing the *AIC*, and the values of *d* and *D* through the unit root test. A step-by-step approach that allows to limit the search field, reducing the effort of calculation and the total time needed (in any case, the process may take several minutes). In the example this method finds and adjusts an ARIMA model (0,0,1) (1,1,0), which is considered as optimal.

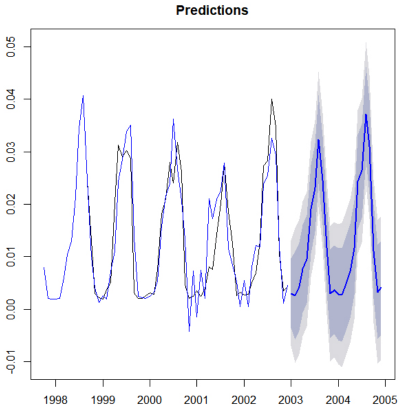


```
ARIMA(0,0,1)(1,1,0)[12]

Coefficients:
         ma1     sar1
      0.7959  -0.6327
s.e.  0.1165   0.1056

sigma^2 estimated as 2.637e-05:  log likelihood=193.9
AIC=-381.8   AICc=-381.28   BIC=-376
```

Prediction of future values

This function calculates the predictions for the number of periods *h* desired, which can be changed by accessing the function forecast with the argument *FORECAST* and by changing the value of *h*. These predictions are shown in figure XV1.2 with confidence intervals. Confidence bands can be colored by specifying *shaded=TRUE* argument within the *FORECAST*, or simply drawing lines. The argument *LINES* allows to modify the adjusted values, or model predictions for all observed periods.

Figure XV1.2. Predictions of the ARIMA model.

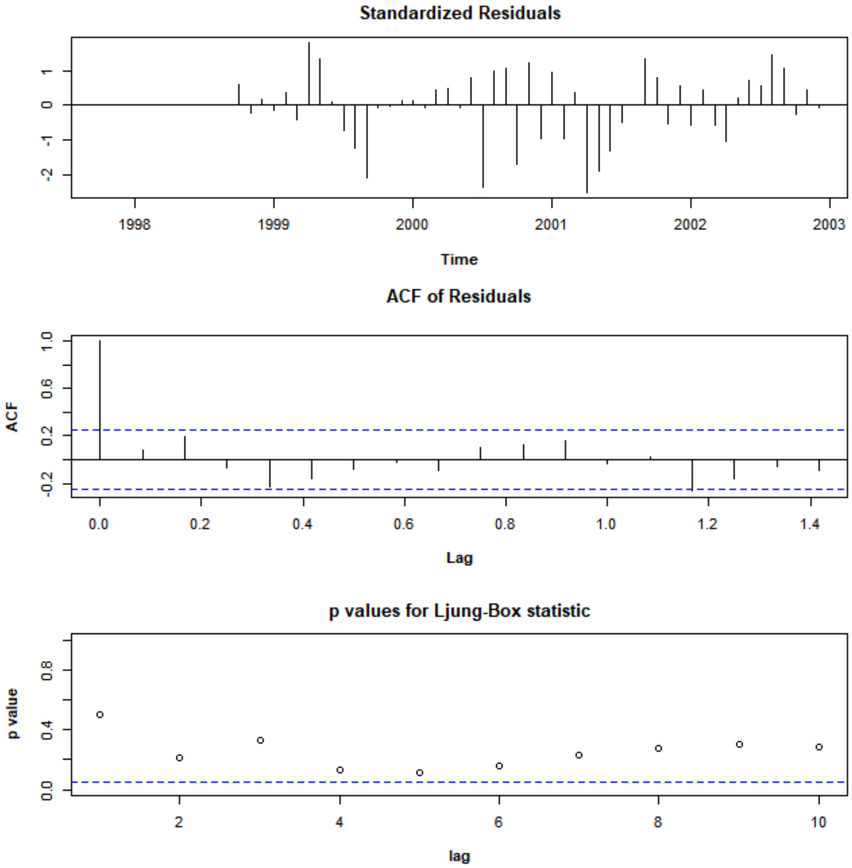


Model diagnostics

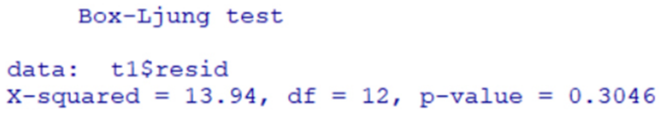
It is advisable to check that the chosen model is complete and new elements should not be added. For this, the analysis of residuals is started with a graph (Figure XV1.3). There, it is possible to check how the residuals seem to behave randomly, without predictable sequences or a structure of autocorrelations, that is to say, in accordance with the model referenced above as "white noise". The *graph* ACF should not show significant autocorrelations (remembering that the first, always equal to 1, is not taken into account), as in this example.

Noting the *p*-values corresponding to the Ljung-boxs test, the assumption that autocorrelations estimated with zero mean and variance $\frac{1}{n}$ are normal, is not rejected for all the delays. All or almost all of the p values must be above the dotted blue line of 5%, as in this example. Therefore, it seems reasonable to conclude that residuals behave like white noise, and the model is adequate.

Figure XV1.3. Residual graphs, *FAS* of residuals,and *p*
values of the Ljung-Box test for every delay of the automated ARIMA model.

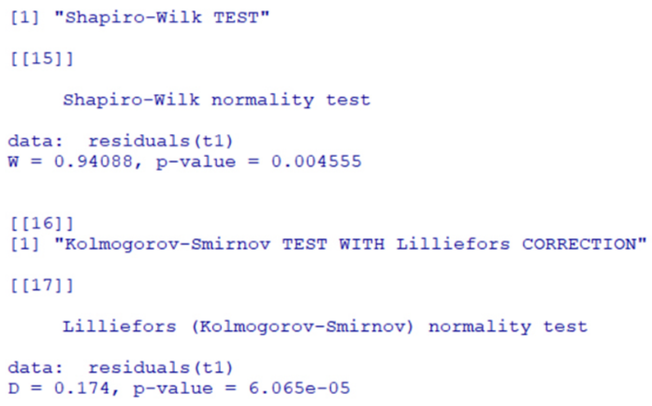


The Box-Ljung test, as shown below, performs a test set to contrast the null hypothesis that it is white noise. The p-value = 0.305 higher than the usual significance level of 0.05 indicates that you can accept that this is white noise and, therefore, the residuals do not retain a structure that must be explained by a more complex model.



```
          Box-Ljung test

data:  t1$resid
X-squared = 13.94, df = 12, p-value = 0.3046
```

It is also possible to check the normal distribution of the residuals, applying the Shapiro-Wilk test and the Kolmogorov-Smirnov with the correction of Lilliefors to residuals. The residuals do not have a Normal distribution, since both the Shapiro-Wilk test as in the Kolmogorov-Smirnov p < 0.05 .



```
[1] "Shapiro-Wilk TEST"

[[15]]

     Shapiro-Wilk normality test

data:  residuals(t1)
W = 0.94088, p-value = 0.004555


[[16]]
[1] "Kolmogorov-Smirnov TEST WITH Lilliefors CORRECTION"

[[17]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(t1)
D = 0.174, p-value = 6.065e-05
```

Deseasonalization

In series with marked seasonality, such as this example, it is difficult to appreciate the variations hidden or masked behind the seasonality and, therefore, to find out the changes that occur. For this reason, seasonal variations can be analyzed in more detail and seasonality of the series eliminated to study the rest.

Figure XV1.4. Monthly values of seasonality in the series.

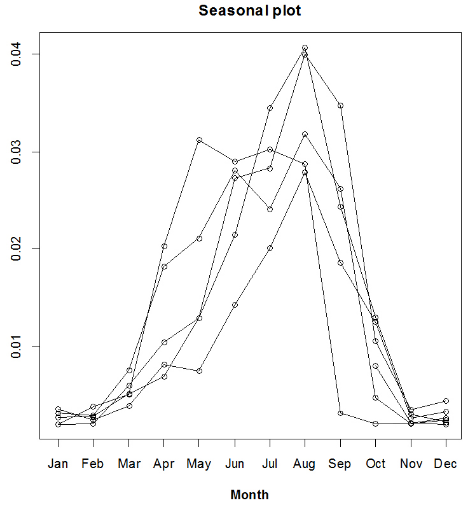
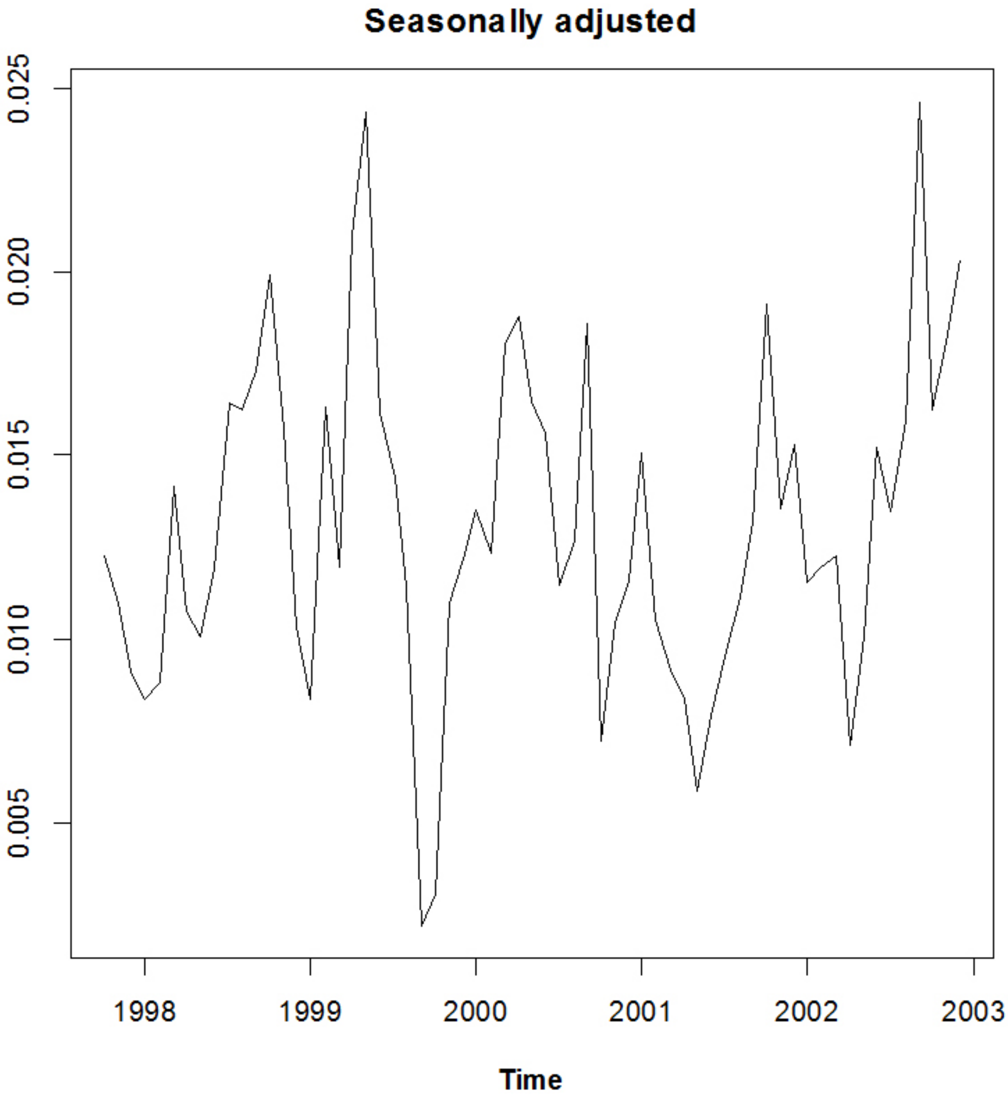


Figure XV1.4 shows the values of monthly seasonality. As can be clearly seen, highly seasonal values are those corresponding to the month of August and next to this month, and the low seasonality occurs from November to February, all of them at similar levels.

Figure XV1.5 shows the seasonally adjusted series, where changes in the series discounting the effect of seasonality can be observed.

Figure XV1.5. Seasonally adjusted series.



MANUAL ARIMA MODEL WITHOUT INDEPENDENT VARIABLES

Instead of using the Expert Modeler, -from autocorrelation plots or knowledge of the series- a particular model can be determined, especially in this case that the normality of residuals is not complied.

This requires defining the arguments *order* (non-seasonal component of the model) and *seasonal* (seasonal component of the model). Both arguments are a vector with three numbers (p, d, q) that correspond to the auto regressive component (p), integration component (d), and moving average component (q).

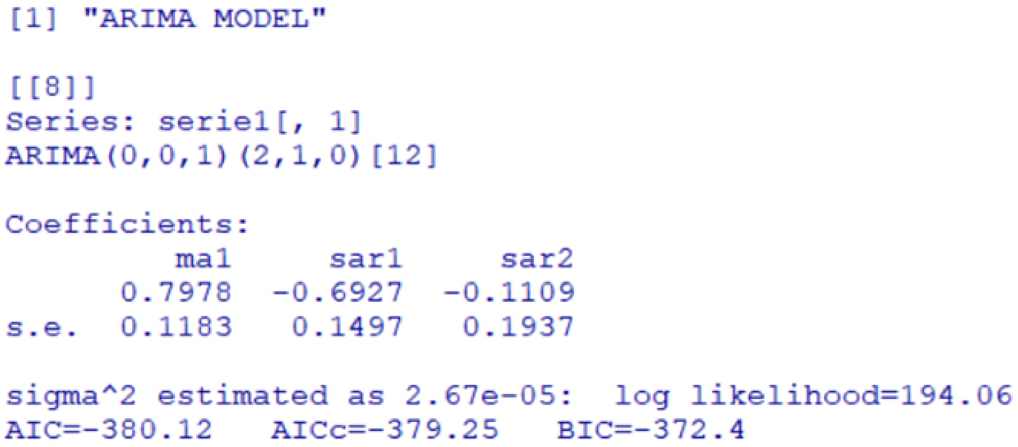
The automated ARIMA model selected (0,0,1) parameters for non-seasonal component and (1,1,0) for the seasonal component. In the manual model we chose the following parameters: *order=c(0,0,1)* and *seasonal=c(2,1,0)*. That is to say, the only change is that the value of p was changed in the seasonal component, which was 1 to a value of 2. Therefore, we believe that the autocorrelation occurs not only with the previous month, but a month is also correlated with the value of two months ago.

It makes no sense to change the integration component (*d*) in the manual model, since the series was stationary and, therefore, the value zero is the most appropriate.

It would be possible to make tests and also change the component *q*, if it is considered that there is innovation in the series, that is to say, that there are a few changes that do not respond to the inertia or the previous historical behavior, but that are new in each period. As already explained in the introduction, it is always a good idea that the model be as simple as possible, i.e., that the values of *p*, *d* and *q* be smaller

The autocorrelation and if the series is stationary or non-stationary are the same results that were automated with the model.

To see if this model is better than that obtained with the expert system, the estimated variance of residuals *sigma^2 estimated* should be small, and the value of *AIC* is best smaller (or more negative). In the case of the automated model the lowest AIC value was -381.8. In the results shown below of the manual model is noted that the value of *AIC* is -380.1. Therefore, the model obtained by the manual system is the best.



```
[1] "ARIMA MODEL"

[[8]]
Series: serie1[, 1]
ARIMA(0,0,1)(2,1,0)[12]

Coefficients:
         ma1     sar1     sar2
      0.7978  -0.6927  -0.1109
s.e.  0.1183   0.1497   0.1937

sigma^2 estimated as 2.67e-05:  log likelihood=194.06
AIC=-380.12   AICc=-379.25   BIC=-372.4
```

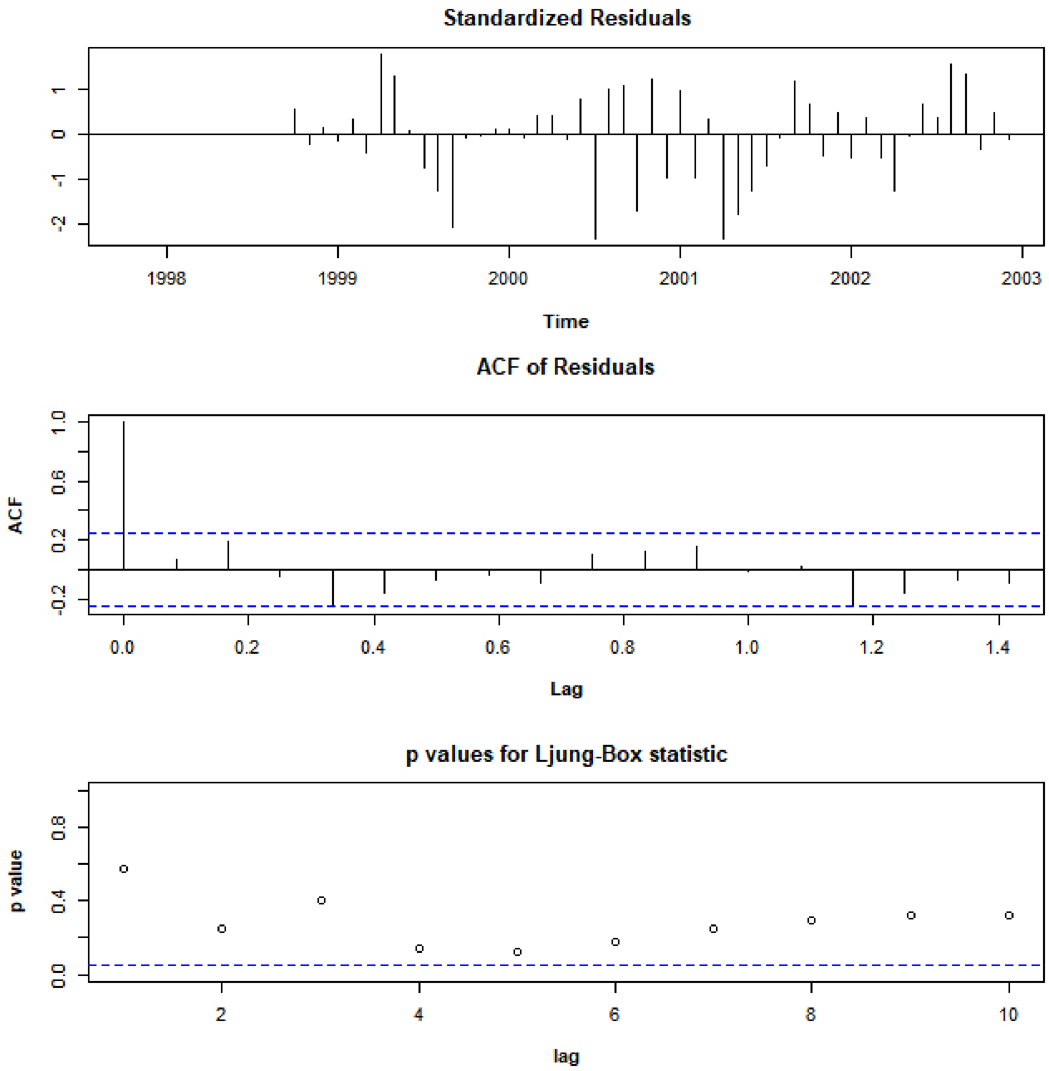
Model Diagnostics

As before, it is necessary to check that the chosen model is complete and new elements should not be added. To do this, the analysis of residuals with a graph (figure XV1.6) starts again.

There, it is possible to check how the residuals seem to behave randomly, without predictable sequences or a structure of autocorrelations, that is to say, in accordance with the model referenced above as "white noise". The graph *ACF* should not show significant autocorrelations (remembering that the first, always equal to 1, is not taken into account), as in this example.

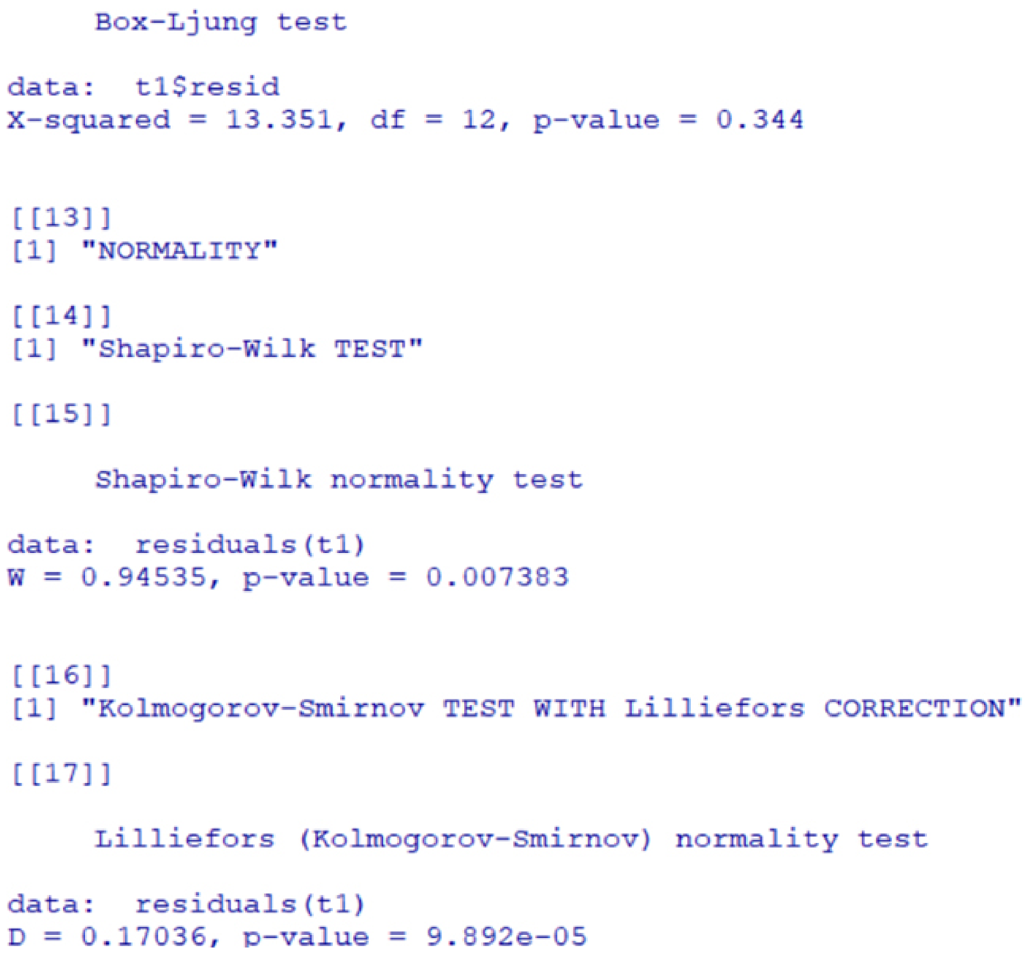
Noting the *p* -values corresponding to the Ljung-boxs test, the assumption that autocorrelations estimated with zero mean and variance $\frac{1}{n}$ are normal is not rejected for all the delays. All or almost all of the *p* values must be above the dotted blue line of 5%, as in this example.

Figure XV1.6. Residual graphs, *FAS* of residuals, and *p* values of the Ljung-Box test for every delay of the manual ARIMA model.



The Box-Ljung test, as shown below, performs a test set to contrast the null hypothesis that it is white noise. The p-value = 0.344 higher than the usual significance level of 0.05 indicates that there may be accepted that this is white noise and, therefore, the residuals do not retain a structure that must be explained by a more complex model.

Unlike the automated model, in this case the residuals have a Normal distribution, since both the Shapiro-Wilk test as in the Kolmogorov-Smirnov $p > 0.05$.



```
     Box-Ljung test

data:  t1$resid
X-squared = 13.351, df = 12, p-value = 0.344


[[13]]
[1] "NORMALITY"

[[14]]
[1] "Shapiro-Wilk TEST"

[[15]]

     Shapiro-Wilk normality test

data:  residuals(t1)
W = 0.94535, p-value = 0.007383


[[16]]
[1] "Kolmogorov-Smirnov TEST WITH Lilliefors CORRECTION"

[[17]]

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(t1)
D = 0.17036, p-value = 9.892e-05
```

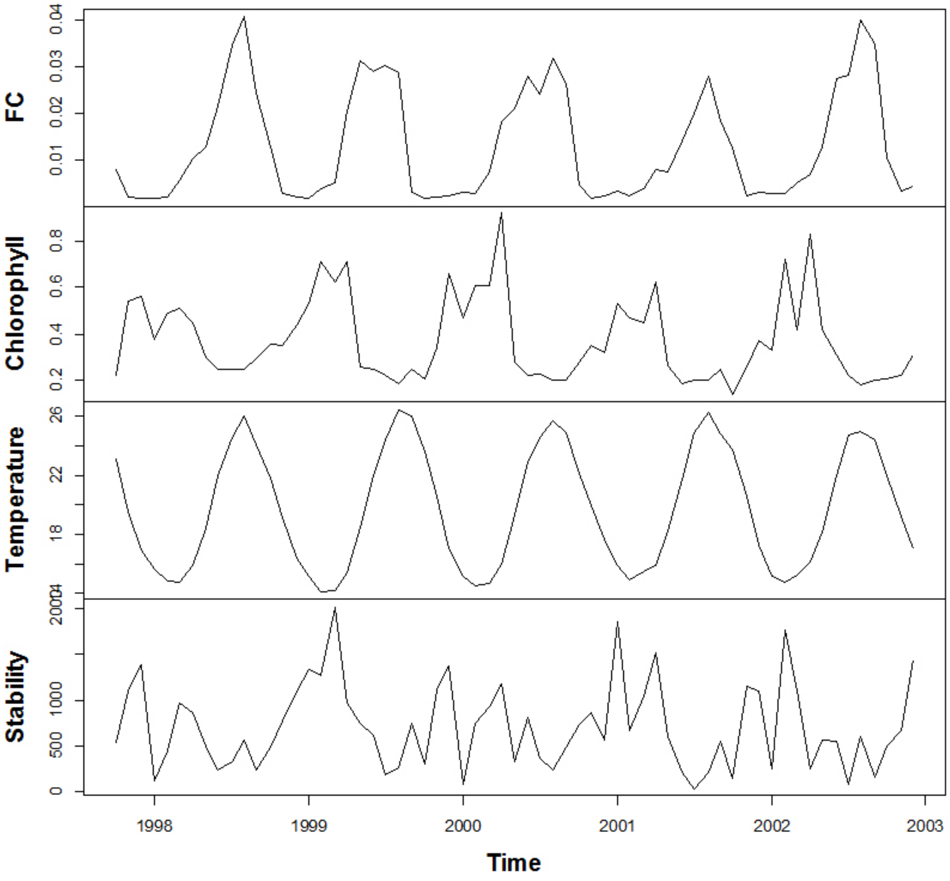
ARIMA MANUAL MODEL WITH INDEPENDENT VARIABLES

Once our time series has already been analyzed, you can now determine the effect of variables on it. In our example, the objective was to determine the effect of the average monthly values of temperature, chlorophyll, and stability of the water column on the *FC* of the anchovy.

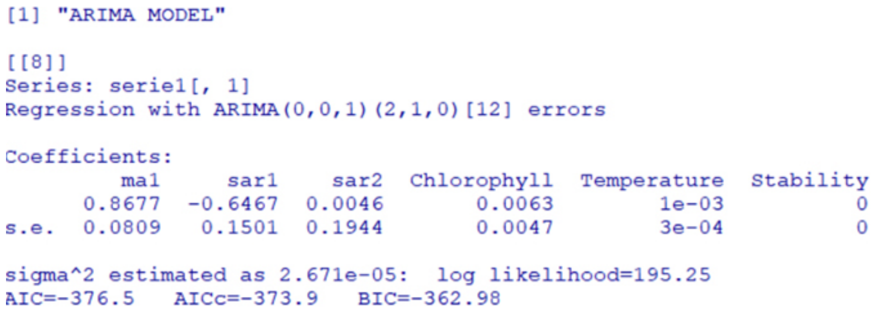
To include independent variables in the model, simply specify them in the argument *varY*. These variables can be included in both the automated model as in the manual. Since the manual model was a bit better than the automated, the example will be made including the independent variables with the manual model.

When independent variables are included in the model, a graph it comes out as a result of those shown above (Figure XV1.7), which shows the temporal evolution of the series (condition factor of the anchovy, *FC*) and also of the independent variables (chlorophyll, temperature and stability).

Figure XV1.7. Temporary representation of the condition factor of the anchovy, stability, temperature, and chlorophyll, in the Strait of Sicily.



The value of AIC with the covariates is -376.5 , while without them was (-380.1). The difference is so small that it can be stated that the covariates do not add useful information and, therefore, the *FC* of the anchovy does not depend on the temperature, chlorophyll, and stability in the water column.



```
[1] "ARIMA MODEL"

[[8]]
Series: serie1[, 1]
Regression with ARIMA(0,0,1)(2,1,0)[12] errors

Coefficients:
         ma1     sar1    sar2  Chlorophyll  Temperature  Stability
      0.8677  -0.6467  0.0046       0.0063        1e-03          0
s.e.  0.0809   0.1501  0.1944       0.0047        3e-04          0

sigma^2 estimated as 2.671e-05:  log likelihood=195.25
AIC=-376.5   AICc=-373.9   BIC=-362.98
```

Value

The following graphs are displayed: 1) changes in variables over time (only if independent variables are included), 2) residuals, the simple autocorrelation function (*ACF*) and the partial correlation coefficient (*PACF*), 3) the model predictions, 4) Ljung-Box test for each delay of the model, 5) the seasonal chart and 6) the graph seasonally adjusted, the last two are displayed if the argument *frequency* is greater than 1.

References

Basilone, G., Guisande, C., Patti, B., Mazzola, S., Cuttitta, A., Bonanno, A., Vergara, A.R. & Maneiro, I. (2006) Effect of habitat conditions on reproduction of the European anchovy (*Engraulis encrasicolus*) in the Strait of Sicily. *Fisheries Oceanography*, 15: 271-280.

Box, G.E.P. & Jenkins, J.M. (1976) *Time series analysis: Forecasting and control*. Holden-Day,San Francisco, CA.

Dickey, D.A. & Fuller, W.A. (1979) Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74, 427-431.

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Hyndman, R.J. & Khandakar, Y. (2008) Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, 27: 1-22.

Hyndman, R.J., Athanasopoulos, G.,Razbash, S., Schmidt, D., Zhou, Z., Khan, Y., Bergmeir, C. & Wang, E. (2018) Forecasting functions for time series and linear models. R package version 8.3. Available at: http://CRAN.R-project.org/package=forecast.

Trapletti, A & Hornik, K. (2017) tseries: Time Series Analysis and Computational Finance. R package version 0.10-38. Available at: https://CRAN.R-project.org/package=tseries.

Examples

```
## Not run:

data(ZXV1)

#Automated Arima Model without independent variables

XV1(data=ZXV1, varY="FC", frequency=12, start=c(1997,10))

#ARIMA manual model without independent variables

XV1(data=ZXV1, varY="FC", frequency=12, start=c(1997,10), order=c(0,0,1),
seasonal=c(2,1,0))

#ARIMA manual model with independent variables

XV1(data=ZXV1, varY="FC", frequency=12, start=c(1997,10), varX=c("Chlorophyll",
"Temperature", "Stability"), order=c(0,0,1), seasonal=c(2,1,0))

## End(Not run)
```

XV2 *CYCLES*

Description

Identification of cycles and analysis of the relationship between cycles.

Usage

```
XV2(data, varY, varX=NULL, frequency, start, ResetPAR=TRUE, PAR=NULL,
TS=NULL, PLOTSERIE=NULL, SPECPGRAM=NULL, SPECAR=NULL, COHERENCE=NULL,
SPECTRUM=NULL, PLOTSPECTRUM=NULL, file="Output.txt")
```

Arguments

`data` Data file.

`varY` Dependent variable.

`varX` Independent variable.

`frequency` Number of observations per unit of time.

`start` The date of the first observation. It can either be a single number or a vector of two integers, specifying a natural unit of time.

`ResetPAR` If it is FALSE, the default condition of the function PAR are not placed and those defined by the user on previous graphics are maintained.

`PAR` It accesses the function PAR which allows to modify many different aspects of the graph.

`TS` It accesses the function ts that allows to modify the parameters of the time series.

`PLOTSERIE` It accesses the function that allows to modify the first graph with the time series.

`SPECPGRAM` It accesses the function spec.pgram which allows to modify the periodogram of the dependent variable.

`SPECAR` It accesses the function spec.ar that allows to modify the density plot of the dependent variable.

`COHERENCE` It accesses the function coh which allows to modify the graph of coherence between the dependent and independent variable.

`SPECTRUM` It accesses the function spectrum which calculates the time-lag between the dependent and independent variable.

`PLOTSPECTRUM` It accesses the function which allows to modify the graph that shows the lag between the dependent variable and the independent.

`file` TXT FILE. Output file name with the results of the analysis of the cycles.

Details

XV1. TIME SERIES

XV1.2 CYCLES

The cycle is a common component in the time series (Guisande et al., 2011). Its typical form is sinusoidal, with a phase of growth to reach the maximum value, followed by a declining phase until reaching its minimum value, repeating this scheme indefinitely. Unlike the seasonal variations, it is not always regular, and the length of the different stages is often changing, which makes it difficult to perform its analysis. This is present in virtually all economic series, without knowing the reasons for its occurrence or the root causes of the existence of cycles.

Spectral analysis is a statistical procedure that can be used to identify periodic behavior in a time series, and to study the existence of cycles or regular variations whose length (period) is usually higher than a year. The process breaks down the set of all observations on the series in recurring (components) elements of different frequencies, and allows to identify the most relevant ones. It is particularly useful for detecting complex cyclical structures, i.e., when there are several cycles of different overlapping cycles, since when there is a single cyclical pattern, a graphical representation (sequential chart) may be sufficient to identify this.

FUNCTIONS

The time series is performed with the function ts, the function spec.pgram is used to obtain the periodogram of the dependent variable, the function spec.ar is used to represent the graph of density of the dependent variable and the time lag between the dependent variable and the independent is obtained with the function spectrum, all of them from the base package stats.

The graph of consistency between the dependent and independent variable is performed with the function coh of the package seewave (Sueur et al., 2014).

The estimation of missing values is performed with the function na.approx of the zoo package (Zeileis et al., 2014).

EXAMPLE

The angle between the sun and the moon is one of the main causes of the changes that occur in the tides. When the sun and moon are aligned, i.e., the angle is 0 (full and new moon), differences in height between high and low tides are highest. By contrast, when the sun and moon are at an angle of 90° (crescent and gibbous) differences in height between high and low tide are minimal.

The data used are the angles between the sun and moon day, and the difference in meters between daily high and low tide, for 42º8´ North and 14º1´ West from 1/1/2004 to 14/3/2004.

The first objective is to determine the tidal cycle. The second objective is to demonstrate the effect of the angle between the sun and moon on the tides.

The function estimates the missing values, if any, but it is necessary that the row with the date exists. In this example there is a missing date, specifically on 3/1/2004.

Figure XV2.1 shows a clear cyclical variation in both time series, with 5 oscillations in a period of 74 days, which corresponds to a period of approximately 15 days.

We can delve into a spectral analysis, which looks at the behavior or the response of the series to each different frequency. In the periodogram shown in the Figure XV2.2, it can be clearly observed a peak that stands out over the rest of values. This indicates the maximum frequency of the cycle that we have previously observed in the graph of sequence. If there were more overlapping cycles, we would find other peaks in the spectrogram showing the frequencies associated with them.

Figure XV2.1. Temporal evolution of the angle between the sun and the moon, and the difference between high and low tide.

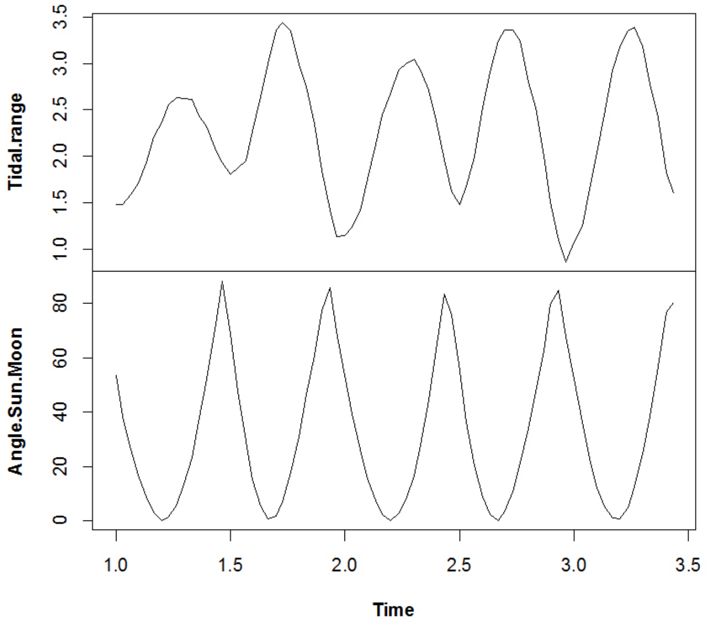
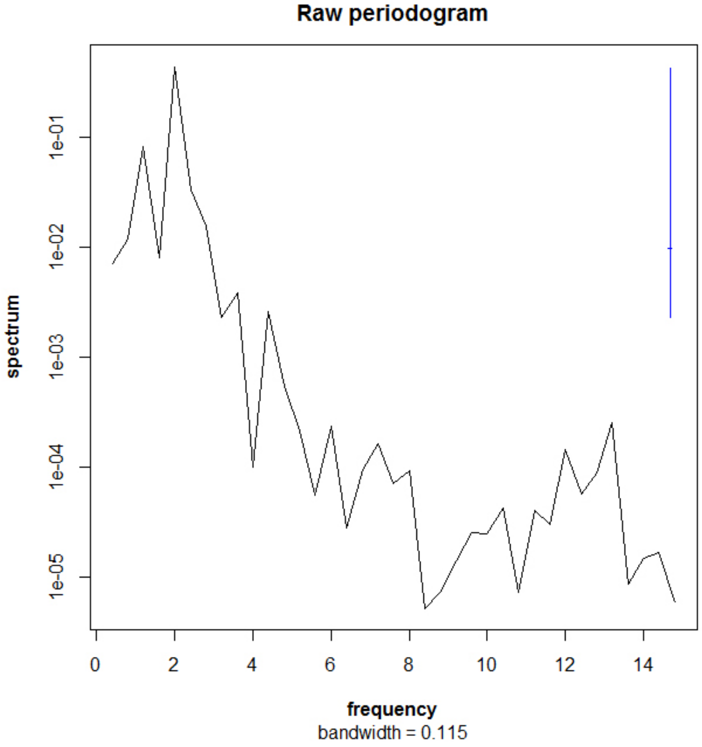
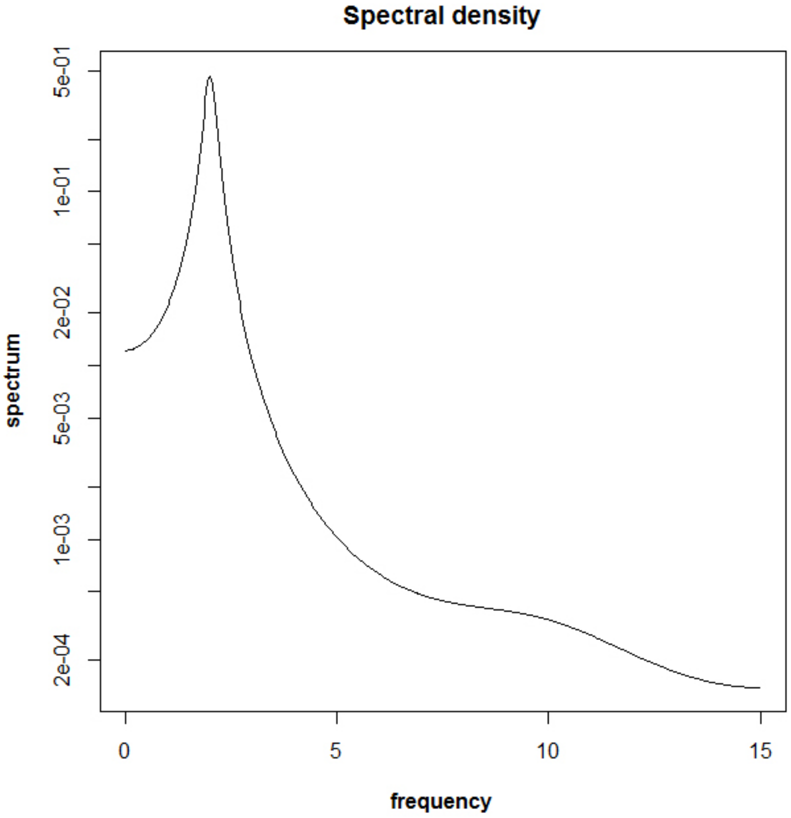


Figure XV2.2. Periodogram of the time series of the difference between high and low tide.



The previous periodogram can be represented by the spectral density plot (Figure XV2.3), which is simply a smoothed version of the periodogram, eliminated by filtering the noise formed by the random variations, or generally any high frequency variation. The effect of smoothing the data allows to see more clearly the peak observed in the periodogram.

Figure XV2.3. Spectral density plot.



The results show that using the first graph (*Raw periodogram*) the period is 15 days, whereas the second graph (*spectral density*) is 14.89 days. It may be more convenient to use the first graph when many missing values in the series, i.e., the periodogram.

The second objective is to find out if the difference between high tide and low tide is related to the angle between the sun and the moon. For that reason, we must build the cross-spectrum for both variables. The cross spectrum usually is interpreted on the basis of some processed variables. One of the most important is the coherence square, represented in the diagram below, which takes values between 0 and 1. A value close to 0 indicates that both variables are not related, and close to 1 indicates a near-linear relationship between the two. The horizontal scale shows the frequency in khz, since it is the most useful unit for the study of signals in the field of electronics in which this technique is usually applied. The values should be multiplied by a thousand to interpret them directly as a fraction of the time unit used.

As figure XV2.4 shows, there is a value close to 1 that stands out clearly, showing the frequency corresponding to a sharp cyclical relationship between two variables. The results show that a consistency of 14.4 days is obtained, which would be the cycle shared between the two time series.

Figure XV2.4. Graph of coherence between the angle that the sun the moon and the tidal cycle form.

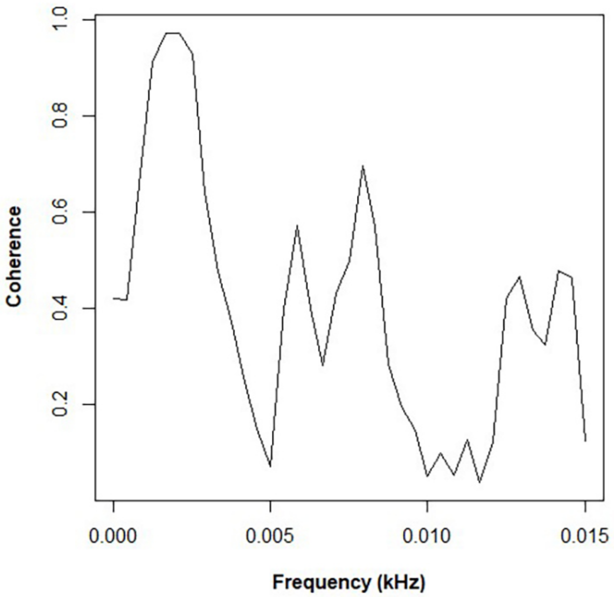
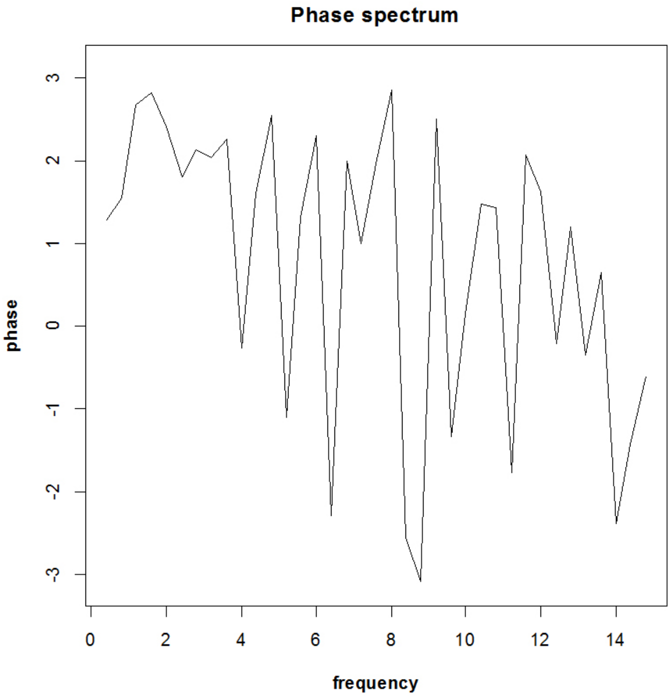


Figure XV2.5. Phase diagram of the angle between the sun and the moon, and the tidal cycle.



Even with a cycle coincident in time duration for the two variables, it is possible that there is a lag or delay between the two. To find out, we build the phase diagram shown in Figure XV2.5. In the diagram there are irregular oscillations - random occurrence - about zero, with amplitude between -3 and 3 days, for the majority of the frequencies, but for the matching frequency that we have detected, below the frequency 0.1 , there are several values that seem to stabilize around -2.

The lag that is obtained (*lag*) is 1.8, which means that the effect of the angle between the sun and the moon on the tides occurs with a delay of 1.8 days.

Value

The following graphs are represented: 1) changes in variables over time, 2) periodogram of the dependent variable, 3) spectral density plot of the dependent variable, 4) graphic coherence between the dependent and independent variable and 5) diagram phase between the dependent and independent variable. The last two graphs are displayed if there is an independent variable.

References

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, STATISTICA y SPSS*. Ediciones Díaz de Santos, Madrid, 978 pp.

Sueur, J., Aubin, T., Simonis, C., Brown, E.C., Desjonqueres, C., Gasc, A., LaZerte, S., Lees, J., Lellouch, L., Pavoine, S., Villanueva-Rivera, L.J., Ross, Z. & Witthoft, C.G. (2014) Sound analysis and synthesis. R package version 1.7.6. Available at: `http://CRAN.R-project.org/package=seewave`.

Zeileis, A., Grothendieck, G., Ryan, J.A. & Andrews, F. (2014) Sound analysis and synthesis. R package version 1.7-11. Available at: `http://CRAN.R-project.org/package=zoo`.

Examples

```
## Not run:

data(ZXV2)

XV2(data=ZXV2, varY=”Tidal.range”, varX=”Angle.Sun.Moon”, frequency=30, start=c(1,1))

## End(Not run)
```

XV3 *PREDICTIVE POWER OF A BINOMIAL LOGISTIC REGRESSION*

Description

It allows to evaluate the predictive power of a binomial logistic regression.

Usage

```
XV3(data, cat, var, resp, start=NULL, nsubsets=20, ResetPAR=TRUE, PAR=NULL,
PLOT=NULL, YLAB=NULL, XLAB=NULL, CEXPCH=1, COLABLINE=”black”, COLOR=”black”, PCH=15,
file1=”Test.csv”, file2=”Coefficients.csv”, na=”NA”, dec=”,”, row.names=FALSE)
```

Arguments

`data` Data file.

`cat` Dependent variable.

`var` Independent variables.

`resp` Category that will be the reference for the dependent variable.

`start` Number of data randomly selected at the beginning.

`nsubsets` Number of random subsets.

`ResetPAR` If it is FALSE, the default condition of the function PAR is not placed and maintained those defined by the user in previous graphics.

`PAR` It accesses the function PAR which allows to modify many different aspects of the graph.

`PLOT` It accesses the function plot.default which allows to modify the graph of the regression.

`YLAB` Legend of the axis Y.

`XLAB` Legend of the axis X.

`CEXPCH` Size of the graphic symbols.

`COLABLINE` Color of the line of the regression model.

`COLOR` Color symbols.

`PCH` Type of symbol.

`file1` CSV FILE. Filename with the percentage of cases correctly identified for each random subset.

`file2` CSV FILE. Filename with the coefficients of the regression models.

`na` CSV FILES. Text that is used in the cells without data.

`dec` CSV FILES. It defines if the comma "," is used as decimal separator or the dot ".".

`row.names` CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

Once a logistic regression model is obtained, which in principle using a series of qualitative and quantitative variables, it may be valid for predicting to what category a case belongs, the idea now is to be able to evaluate the predictive power of the model.

The algorithm consists of taking a number of random samples from the data set, elaborating the logistic regression model with this data and estimating the percentage of cases correctly identified considering the total data set.

Finally, we apply the four accumulation regressions described in the function XI2, to relate the number of data chosen randomly in the subsample and the percentage of cases correctly identified.

A logistic regression would have a high predictive power, if a percentage of success close to 100% is achieved using half of the data.

FUNCTIONS

It uses the function lillie.test of the package (Gross, 2013) to perform the test Kolmogorov-Smirnov normality with Lilliefors' correction and the function nls of the base package stats for the estimation of the models.

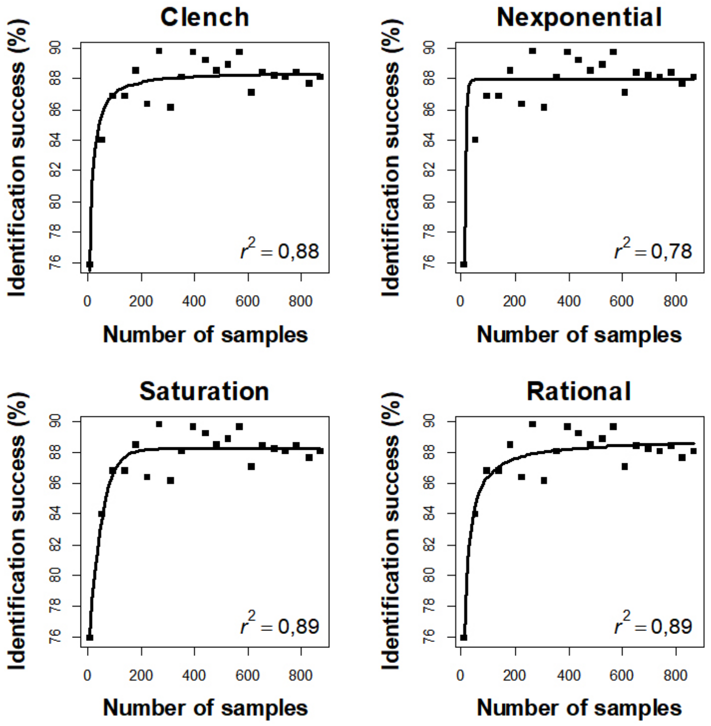
To estimate binomial logistic regression, the glm function of the package stats is used. To estimate the model by steps, the stepAIC function of the package MASS (Venables & Ripley, 2002; Ripley et al., 2014) was used.

EXAMPLE

The study was to examine patients who have been diagnosed with a type of cancer, as well as a sample of people who do not. The objective is to determine if variables such as gender, age, marital status (unmarried, married and separated) and presence/absence of blood in the urine, can be an indicator of the presence of cancer.

As samples are chosen randomly, each time the function is executed, the logistic regression models that are estimated can vary and, therefore, also the accumulation curves.

The graph shows that the best adjustments are given by the Rational, Clench and Saturation curves. When half of the data is used to construct the logistic regression model, the asymptote is already reached and the model predicts about 88% of cases correctly. Therefore, the prediction power of the model has been clearly identified.



Value

A TXT file is obtained, which is called "Output.TXT" with the equations of the four accumulation regressions. A CSV file with the coefficients of the accumulation regressions. A graph with the four accumulation regressions relating the number of randomly selected data and the percentage of cases correctly identified. A CSV file is also obtained, which is called "Model predictions.CSV" with the predictions for each case, resulting from applying the logistic regression to the last random subsample. Finally, the replies window shows the percentage of cases correctly identified that each accumulation regression predicts in the asymptote.

References

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: http://CRAN.R-project.org/package=MASS.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXI7)
```

```
XV3(data=ZXI7, cat="Cancer" , var=c("Age", "Gender", "Status", "Blood"),
gruporesp="Si", start=10)

## End(Not run)
```

XV4 *PREDICTIVE POWER OF A MULTINOMIAL LOGISTIC REGRESSION*

Description

It allows to evaluate the predictive power of a multinomial logistic regression.

Usage

```
XV4(data, cat, var, start=NULL, nsubsets=20, ResetPAR=TRUE, PAR=NULL,
PLOT=NULL, YLAB=NULL, XLAB=NULL, CEXPCH=1, COLABLINE="black", COLOR="black", PCH=15,
file1="Test.csv", file2="Coefficients.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Data file.
`cat`	Dependent variable.
`var`	Independent variables.
`start`	Number of data randomly selected at the beginning.
`nsubsets`	Number of random subsets.
`ResetPAR`	If it is FALSE, the default condition of the function PAR is not placed and maintained those defined by the user in previous graphics.
`PAR`	It accesses the function PAR which allows to modify many different aspects of the graph.
`PLOT`	It accesses the function plot.default which allows to modify the graph of the regression.
`YLAB`	Legend of the axis Y.
`XLAB`	Legend of the axis X.
`CEXPCH`	Size of the graphic symbols.
`COLABLINE`	Color of the line of the regression model.
`COLOR`	Colour symbols.
`PCH`	Type of symbol.
`file1`	CSV FILE. Filename with the percentage of cases correctly identified for each random subset.
`file2`	CSV FILE. Filename with the coefficients of the regression models.
`na`	CSV FILES. Text that is used in the cells without data.

`dec`	CSV FILES. It defines if the comma "," is used as decimal separator or the dot ".".
`row.names`	CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

As mentioned in the same section of the function XV3.

FUNCTIONS

It uses the function lillie.test of the package (Gross, 2013) to perform the test Kolmogorov-Smirnov normality with Lilliefors' correction and the function nls of the base package stats for the estimation of the models.

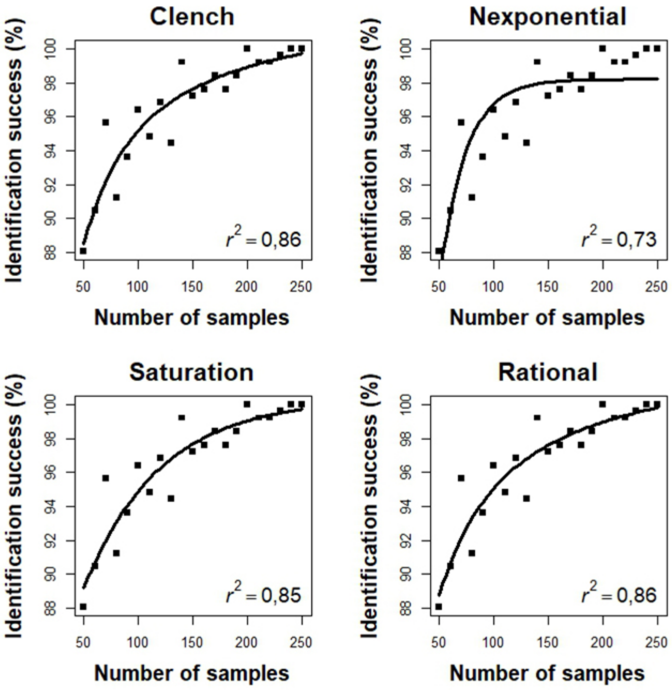
To estimate binomial logistic regression, the glm function of the package stats is used. To estimate the model by steps, the stepAIC function of the package MASS (Venables & Ripley, 2002; Ripley et al., 2014) was used.

EXAMPLE

The data are body measures of several species of fish (continuous variables), as well as the type of mouth they have (categorical variable). The objective is to determine, if it is possible to identify, which family an individual belongs in function of biometrics.

As samples are chosen randomly, each time the function is executed, the logistic regression models that are estimated can vary and, therefore, also the accumulation curves.

The graph shows that the best adjustments are Rational, Clench and Saturation curves. When half of the data is used for the logistic regression model, a value is obtained around 97% and, according to the accumulation curves, the asymptote is around 100%. Therefore, it may be possible that with more data, a 100% of identification success would be achieved with half of the data set.



Value

As mentioned in the same section of the function XV3.

References

Gross, J. (2013) Tests for Normality. R package version 1.0-2. Available at: http://CRAN.R-project.org/package=nortest.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2014) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-33. Available at: http://CRAN.R-project.org/package=MASS.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXI9)

XV4(data=ZXI9, cat="Family", var=c("M2","M6","M8","M9", "M12","M13",
"M15","M16","M17","M21","M22","Mouth"), start=50)

## End(Not run)
```

XV5 *DATA STANDARDIZATION*

Description

Variables are standardized to any range defined by the user (by default between 0 and 1) or you can simply standardize all the variables depending on one of them.

Usage

```
XV5(data, var, varRef=NULL, varCode=NULL, min=0, max=1, file="Standardized.csv",
na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Data file.
`var`	Variables to be standardized.
`varRef`	If a variable is specified, all are standardized according to it, that is, all are divided by that variable.
`varCode`	Optionally, variables of the original matrix can be selected, which are not standardized, and are exported in the output file with standardized variables. For example, this allows to choose variables which are codes of rows.
`min, max`	Minimum and maximum values of the standardization range.
`file`	CSV FILE. Name of the file with the standardized variables.
`na`	CSV FILE. Text that is used in the cells without data.
`dec`	CSV FILE. Defines whether as decimal separator is used the comma "," or the dot ".".
`row.names`	CSV FILE. Logical value which defines whether there are identifiers on the rows or a vector with a text for each of the rows.

Details

EXAMPLE

The data are socioeconomic parameters of 57 countries in Europe, Africa and America. The variables used were male and female life expectancy at birth (in years of life), the mortality rates, infant mortality, birth, and fertility, the gross domestic product per capita (in thousands of dollars per year) and the literacy rate for men and women (in percentage) in the year 2000. The data were obtained from The World Bank (`http://www.worldbank.org/`). All the variables are standardized to the default range of 0 to 1 and the codes of the continent and the country are added.

Value

A CSV file is exported with the standardized variables and with the variables included in the argument *varCode* if any variable was included in this argument.

Examples

```
## Not run:

data(ZXIII2)

XV5(data=ZXIII2, var=c("LifeExpF", "LifeExpM", "Mortinf", "PIB_cap",
"Birthrate", "Mortality", "Fertility", "LiteracyM", "LiteracyF"),
varCode=c("Continent", "Country"))

## End(Not run)
```

XV6 *DENDROGRAM ON A PRINCIPAL COMPONENT ANALYSIS*

Description

A Dendrogram is applied on a Principal Components analysis.

Usage

```
XV6(data, var, labels, cat=NULL, por=80, k=NULL, pthreshold=0.05,
ellipse=FALSE, convex=FALSE, dim=c(1,2), size=c(1,5), showCluster=TRUE,
VIF=FALSE, threshold=10, method="overlap", minimum=TRUE,
ResetPAR=TRUE, PAR=NULL, PCA=NULL, SCATTERPLOT=NULL, HCLUST=NULL, CLUSTER=NULL,
BOXPLOT=NULL, mfrowBOXPLOT=NULL, LabelCat=NULL, COLOR=NULL, COLORC=NULL,
COLORB=NULL, PCH=NULL, XLIM=NULL, YLIM=NULL, XLAB=NULL, YLAB=NULL,
ylabBOXPLOT=NULL, LEGEND=NULL, MTEXT= NULL, TEXTvar=NULL, TEXTlabels=NULL,
arrows=TRUE, larrow=0.7, colArrows="black", quadratic=FALSE, file1="Output.txt",
file2="Cat loadings.csv", file3="Descriptive statistics of clusters.csv",
file4="Original data and cluster number.csv", file5="Var loadings-Linear.csv",
file6="Cat loadings-Linear.csv", file7="Table cross-validation-Linear.csv",
file8="Cases cross-validation-Linear.csv", file9="Table cross-validation-Quadratic.csv",
file10="Cases cross-validation-Quadratic.csv", file11="Plots VARSEDIG.pdf",
file12="U Mann-Whitney test.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

`data` Data file.

`var` Variables that are included in the analysis.

`labels` Variable that allows to display a label for each case.

`cat` Optionally you can specify a variable to show a grouping in the plot of Principal Components.

`por` Cut-off threshold specifying the cumulative variance percentage, to determine how many axes are selected from the Principal Components analysis. By default it is 80%, which means that the axes are selected until reaching an accumulated variance percentage of 80%.

k Number of clusters in which the Dendrogram is divided. If it is NULL, the algorithm select automatically the maximum number of clusters in which the Dendrogram can be divided, which are those groups that are statistically different in at least one variable according to the U Mann-Whitney test.

pthreshold Threshold probability of the U Mann-Whitney test.

ellipse If it is TRUE, the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *cat* are depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*. If it is TRUE, the ellipses of the clusters in the Discriminant analysis and in the polar coordinate plot of the VARSEDIG algorithm are also calculated.

convex If it is TRUE, the convex hull is calculated for each category in the Principal Components analysis, but only if some variable has been selected in the argument *cat*. If TRUE, the convex hull of the clusters is also calculated in the Discriminant analysis and in the polar coordinate plot of the VARSEDIG algorithm.

dim Vector with two values indicating the axes that are shown in the Principal Components analysis.

size Size range of bubbles. Two values: minimum and maximum size.

showCluster If it is TRUE, the number of each cluster is shown in the Dendrogram.

VIF If it is TRUE, the inflation factor of the variance (VIF) is used to select the highly correlated variables and, therefore, not correlated variables are excluded from the Principal Components analysis.

threshold Cut-off value for the VIF.

method Three different methods for prioritizing the variables according to their capacity for discrimination can be used in the VARSEDIG algorithm. If the method is "overlap", a density curve is obtained for each variable and the overlap of the area under the curve between the two groups of the variable *group* is estimated for all variables. Those variables with lower overlap should have better discrimination capacities and, hence, all variables are ordered from lowest to highest overlap; in other words, from the highest to lowest discrimination capacity. If the method is "Monte-Carlo", a Monte-Carlo test is performed comparing all values of group 1 with group 2, and all values of group 2 with 1. The variables are prioritized from the variable with the lowest mean of all p-values (highest discrimination capacity) to the variable with the highest mean of all p-values (lowest discrimination capacity). If the method is "logistic regression", then a binomial logistic regression is calculated and only significant variables are selected for further analyses with the regression performed by steps using the Akaike Information Criterion (AIC).

minimum If it is TRUE, the algorithm is designed to find a significant discrimination between both groups with the minimum possible number of significant variables. Therefore, only the variables with higher discrimination capacity are selected. If it is FALSE, the algorithm selects all significant variables, and not only those with higher discrimination capacity. This argument is only valid with the methods "Monte-Carlo" and "overlap" and it is useful in those cases that discrimination between the groups is difficult and requires to include as many as variables as possible.

`ResetPAR` If it is FALSE, the default condition of the function PAR are not placed and those defined by the user on previous graphics are maintained.

`PAR` It accesses thePAR function that allows to modify many different aspects of the graphs.

`PCA` It accesses the prcomp function of the stats package.

`SCATTERPLOT` It accesses the function scatterplot of the car package.

`HCLUST` You may access the function hclust of the stats package.

`CLUSTER` Access to the function that allows to modify the graphic representation of the Dendrogram.

`BOXPLOT` Allows to specify the characteristics of the boxplot.

`mfrowBOXPLOT` It allows to specify the boxplot panel. It is a vector with two numbers, for example c(2,5) which means that the boxplots are put in 2 rows and 5 columns.

`LabelCat` It allows to specify a vector with the names of the clusters in the boxplots. They must be as many as clusters.

`COLOR` It allows to modify the colours of the graphic in the Principal Components, but they must be as many as different groups have the variable *cat*.

`COLORC` It allows to modify the colours of the clusters in the Dendrogram, but they must be as many as clusters.

`COLORB` It allows to modify the colours of the clusters in the boxplots, but they must be as many as clusters.

`PCH` Vector with the symbols of the Principal Components plot, which must be as many as different groups have the variable *cat*. If it is NULL they are calculated automatically starting with the symbol 15.

`XLIM, YLIM` Vectors with the axes limits *X* and *Y* of the Principal Components plot.

`XLAB, YLAB` Legends of the axes *X* and *Y* of the Principal Components plot.

`ylabBOXPLOT` You can specify a vector with the legends of the axes *Y* of the boxplots. They should be as many as the number of variables.

`LEGEND` It allows to include or to modify a legend in the Principal Components plot.

`MTEXT` It allows to add text in the margins of the Principal Components plot.

`TEXTvar` It allows to modify the labels of the variables in the Principal Components plot.

`TEXTlabels` It allows to modify the labels of the cases in the Principal Components plot.

`arrows` If it is TRUE the arrows are shown in the scatterplot in the Principal Components analysis.

`larrow` It modifies the length of the arrows in the Principal Components plot.

`colArrows` Colours of the arrows in the Principal Components plot.

`quadratic` If TRUE, a Quadratic Discriminant Analysis is performed, in addition to the Linear Discriminant Analysis.

`file1` TXT FILE. Name of the output file with the results.

`file2` CSV FILE. Name of the output file with the coordinates of the cases of the Principal Components plot.

file3 CSV FILE. Name of the output file with the descriptive statistics of each variable for each of the clusters obtained in the Dendrogram.

file4 CSV FILE. Name of the output file with the original data of the variables and the cluster to which each case belongs.

file5 CSV FILE. Name of the output file with the coordinates of the variables in the Linear Discriminant Analysis plot.

file6 CSV FILE. Name of the output file with the coordinates of the categories in the Linear Discriminant Analysis plot.

file7 CSV FILE. Name of the output file with the prediction table using the cross-validation of the Linear Discriminant Analysis.

file8 CSV FILE. Name of the output file with the group to which each case belongs and the prediction of the Discriminant Analysis using the cross-validation of the Linear Discriminant Analysis.

file9 CSV FILE. Name of the output file with the predictions table using the cross-validation of the Quadratic Discriminant Analysis.

file10 CSV FILE. Name of the output file with the group to which each case belongs and the prediction of the Discriminant Analysis using the cross-validation of the Quadratic Discriminant Analysis.

file11 PDF File. Name of the output file with the graphics obtained from the VARSEDIG algorithm.

file12 CSV FILE. Name of the output file with the obtained probabilities of comparing all the variables among all the clusters with the U Mann-Whitney test.

na CSV FILES. Text that is used in the cells without data.

dec CSV FILES. It defines if a comma "," or a dot "." is used as decimal separator.

row.names CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

The aim of this analysis is to determine what statistically different groups are formed by applying a Principal Components analysis.

The first axis in a Principal Components analysis is the linear combination of the original variables that has maximum variance. The second component is the linear combination of the original variables with maximum variance with the added condition that it is independent of the first (orthogonal), and so on, all the main components can be obtained, which, being independent of each other, contain different information. The independence or absence of correlation means that the new variables or components do not share common information. Each main component, therefore, explains the maximum possible residual variability (which has not already been explained above). Therefore, in a Principal Components analysis the cases are differentiated according to the variables that have greater variability. The idea of the analysis is to determine if statistically different groups are formed associated to the variability observed in the variables.

This analysis can be useful to find different groups when you really do not know what they are. For example, find different species using morphometric variables, without really knowing how many potential species there are and to what species each individual belongs. However, it is important to note that only different groups will be detected if the variables that have more variability give rise to different groups. It is possible that a variable does not present a great variability, but it is important for discriminating groups. This type of differentiation based on variables that do not have high variance, would not be detected in this analysis.

To detect the potential groups being formed, a Dendrogram is applied to the scores obtained from the axes that absorb a greater variance. By default, the axes that absorb 80% of the variability are chosen, but this value can be modified by the user.

Subsequently, a Discriminant Analysis is carried out to determine if the clusters that have been generated are well discriminated, that is, to determine the number of correctly identified cases in each cluster.

Next, a U Mann-Whitney test is performed to determine if there are significant differences in the variables between the clusters.

Finally, the algorithm of the VARSEDIG function is applied (see for more details Guisande et al 2016: Guisande, 2018). With this algorithm it is possible to determine if all the cases of each cluster are statistically different from the other clusters.

The idea of this function is to find the largest possible number of clusters with the highest discrimination percentage. To do this the user should perform tests, modifying the cut-off threshold by specifying the cumulative variance percentage to determine how many axes are selected from the Main Components (by default *by=80*) and the variables to be included, eliminating those that are not correlated and are not useful in the Principal Components analysis, as well as those that have little discrimination power in the Discriminant Analysis.

FUNCTIONS

The Principal Components Analysis was performed with the prcomp function of the stats package. The vif function of the usdm package was used for the calculation of VIF (Naimi, 2013; Naimi et al., 2014). To perform the *biplot* graph the scatterplot function of the car package was used (Fox et al., 2018). The arrows are depicted with the function Arrows of the package IDPmisc (Locher & Ruckstuhl, 2014). The convex hull is estimated with the function chull of the package grDevices. KMO test was performed with the function KMO of the package psych (Revelle, 2018). Bartlett's test sphericity was performed with the function bart_spher of the package REdaS (Maier, 2015). The U Mann-Whitney test is performed with the \ emph wilcox.test function of the base stats package. The comparison between clusters with the VARSEDIG algorithm is done with the VARSEDIM function of the VARSEDIG package (Guisande et al., 2016: Guisande, 2018). The Linear Discriminant Analysis was performed with the functions candisc of the candisc package (Friendly, 2007; Friendly & Fox, 2017) and lda of the MASS package (Venables & Ripley, 2002; Ripley et al., 2018). The Quadratic Discriminant Analysis was performed with the function qda of the MASS package (Venables & Ripley, 2002; Ripley et al., 2018). The graph with one dimension in the Discriminant analysis was performed with the function plot.cancor of the candisc package (Friendly, 2007; Friendly & Fox, 2017).

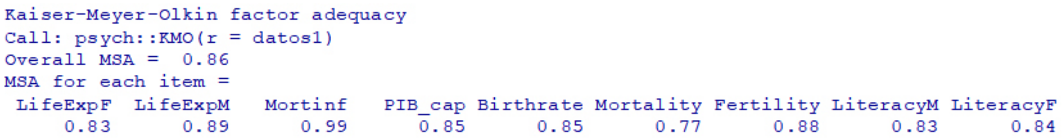
EXAMPLE

The study consisted of analysing the demographic parameters of 57 countries in Europe, Africa and America. The variables used were the male and female life expectancy at birth (in years of life), mortality rates, infant mortality, birth rate, and fertility, the gross domestic product per capita (in thousands of dollars per year) and the rate of literacy of men and women (in percentage) in the year 2000. The data was obtained from the World Bank (`http://datos.bancomundial.org/`). The objective is to see which groups are formed from the Principal Components analysis.

First all variables are used *var=c("LifeExpF", "LifeExpM", "Mortinf", "PIB_cap", "Birthrate", "Mortality", "Fertility", "LiteracyM", "LiteracyF")* and the labels were *labels="Country"*. For purposes only of graphic presentation in the Principal Components, the continent is used as a category *cat="Continent"*. It is important to highlight that the category is not used for any statistical analysis and it is simply used to group the cases with ellipses or with the convex hull in the Principal Components graphic.

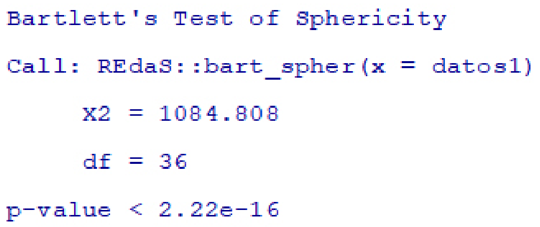
It is always convenient to perform the analysis by eliminating the variables that are not correlated, for which we must specify *VIF=TRUE*. However, it has not been done in the example, because the KMO of all the variables was significant. Therefore, the VIF values are not displayed in the output TXT file.

The first result obtained is the KMO test, which tells us if the variables are adequate for the Principal Components. The value must be greater than 0.5. Therefore, all variables that do not have a value greater than 0.5, could be eliminated from the analysis. In the case that the value is exactly 0.5, it means that it is not possible to estimate the KMO.



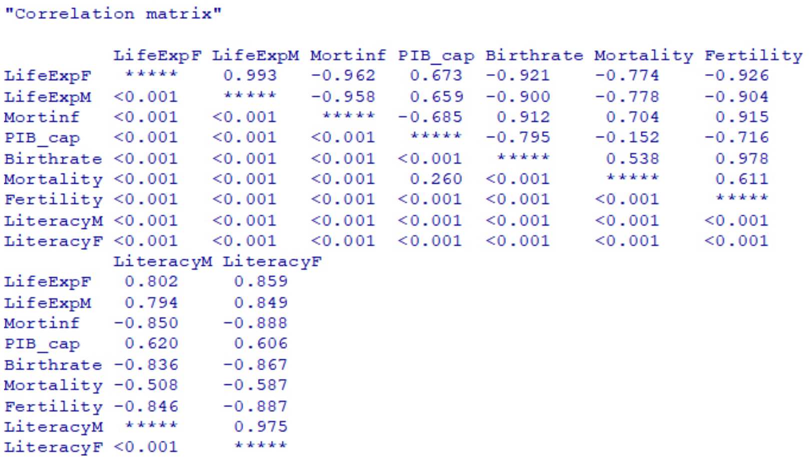
```
Kaiser-Meyer-Olkin factor adequacy
Call: psych::KMO(r = datos1)
Overall MSA =  0.86
MSA for each item =
 LifeExpF  LifeExpM   Mortinf   PIB_cap Birthrate Mortality Fertility LiteracyM LiteracyF
     0.83      0.89      0.99      0.85      0.85      0.77      0.88      0.83      0.84
```

The second statistic that appears is Bartlett's test of sphericity, which tests whether the correlation matrix is an identity matrix, which would indicate that the factor model is inappropriate. A value *p* of the contrast smaller than the level of significance allows rejecting the hypothesis and concluding that there is correlation. Therefore, for the Principal Components analysis to be valid, the probability must be less than 0.05, as it is in this case.



```
Bartlett's Test of Sphericity

Call: REdaS::bart_spher(x = datos1)

     X2 = 1084.808

     df = 36

p-value < 2.22e-16
```

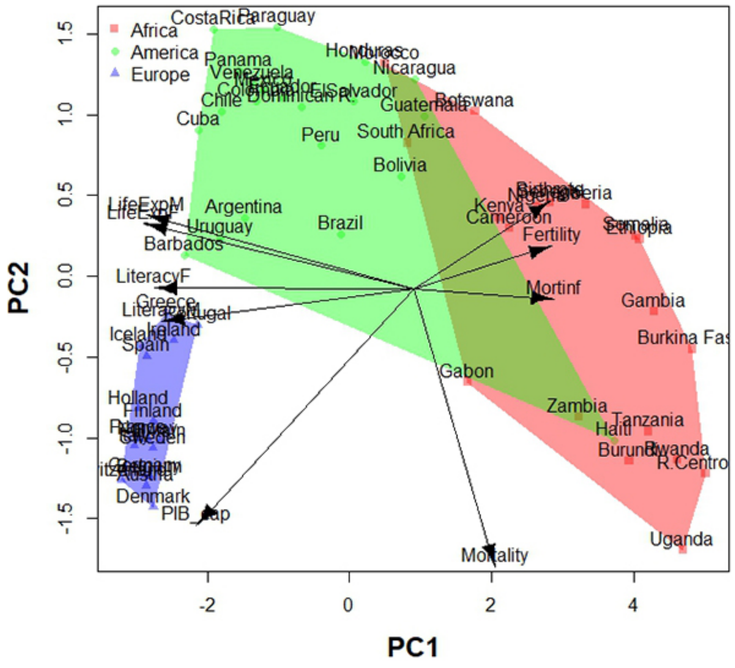
As we can see in the correlation matrix, almost all *p* values are less than 0.05, which indicates that none of the correlation coefficients seems to be zero. Therefore, all the variables are correlated.



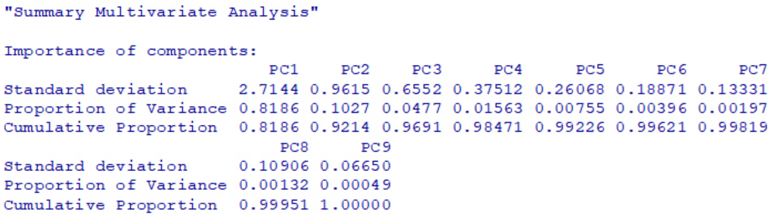
```
"Correlation matrix"

          LifeExpF LifeExpM Mortinf PIB_cap Birthrate Mortality Fertility
LifeExpF   *****    0.993   -0.962   0.673  -0.921    -0.774    -0.926
LifeExpM  <0.001    *****   -0.958   0.659  -0.900    -0.778    -0.904
Mortinf   <0.001   <0.001    *****  -0.685   0.912     0.704     0.915
PIB_cap   <0.001   <0.001   <0.001   *****  -0.795    -0.152    -0.716
Birthrate <0.001   <0.001   <0.001  <0.001   *****     0.538     0.978
Mortality <0.001   <0.001   <0.001   0.260  <0.001     *****     0.611
Fertility <0.001   <0.001   <0.001  <0.001  <0.001    <0.001     *****
LiteracyM <0.001   <0.001   <0.001  <0.001  <0.001    <0.001    <0.001
LiteracyF <0.001   <0.001   <0.001  <0.001  <0.001    <0.001    <0.001
          LiteracyM LiteracyF
LifeExpF   0.802     0.859
LifeExpM   0.794     0.849
Mortinf   -0.850    -0.888
PIB_cap    0.620     0.606
Birthrate -0.836    -0.867
Mortality -0.508    -0.587
Fertility -0.846    -0.887
LiteracyM  *****     0.975
LiteracyF <0.001     *****
```

Figure XV6.1 shows that in the first axis the countries of Europe, America and Africa are perfectly differentiated. Those in Europe are characterized by high life expectancy and high literacy, both in men and women. Those in Africa for their high infant mortality and high birth rate.

Figure XV6.1. Principal Components analysis showing the variability observed in the countries.



The first axis accounts for 81.86%, the second for 10.27% and the third for 4.7% of the variance observed. The first two axes explain 92.14% of the variance. Since the default value of *by=80* was selected, only the first Principal Component axis is selected.



```
"Summary Multivariate Analysis"

Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.7144 0.9615 0.6552 0.37512 0.26068 0.18871 0.13331
Proportion of Variance 0.8186 0.1027 0.0477 0.01563 0.00755 0.00396 0.00197
Cumulative Proportion  0.8186 0.9214 0.9691 0.98471 0.99226 0.99621 0.99819
                           PC8     PC9
Standard deviation     0.10906 0.06650
Proportion of Variance 0.00132 0.00049
Cumulative Proportion  0.99951 1.00000
```

Figure XV6.2 shows the Dendrogram where 4 clusters are grouped, since the argument *k=4* was specified.

Figure XV6.2. Dendrogram with the scores of the axes selected from the Principal Components analysis.

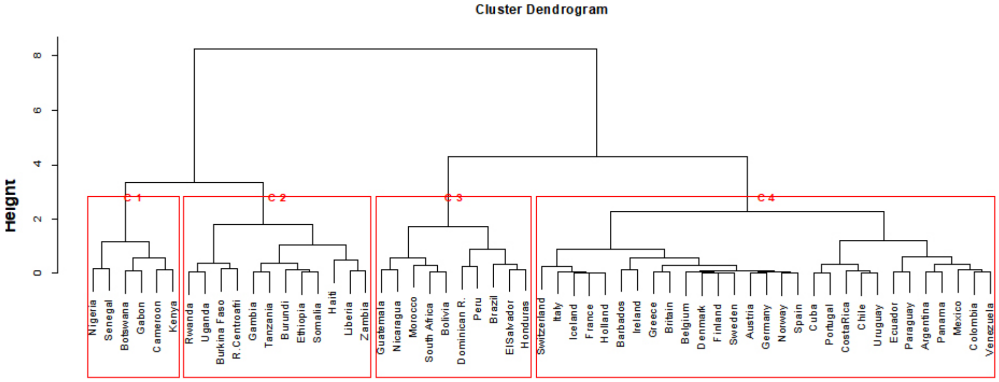
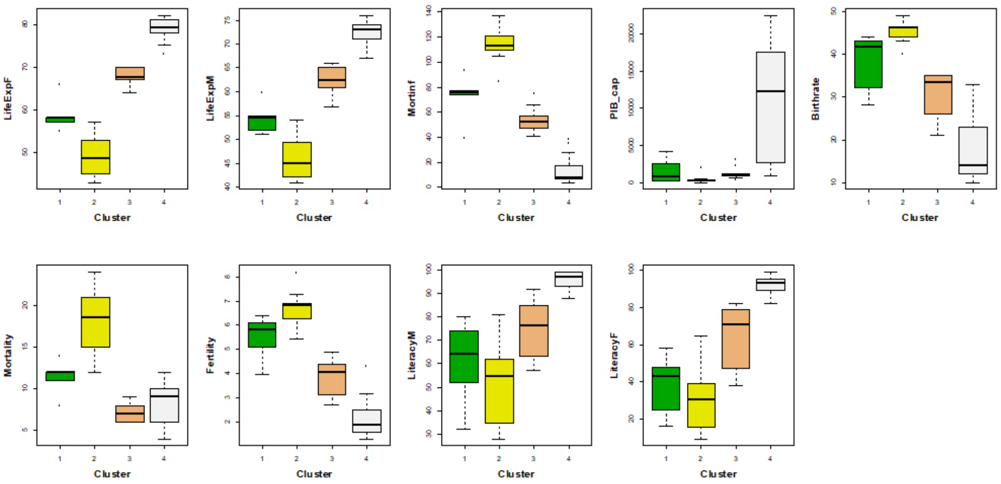


Figure XV6.3 shows the differences between clusters for each of the variables. It is clear the difference in the GDP of Cluster 4, which groups most European countries. Also note the high values of mortality and infant mortality of the group of countries of cluster 2.

Figure XV6.3. Boxplot obtained for each of the variables with the averaged values for each cluster.



The Discriminant Analysis shows that it is possible to correctly discriminate 100% of cases by cross-validation with the Linear method. The first discriminant axis explains most of the variability and discriminates well between the 4 clusters (Figure XV6.4). All the variables seem to be important for the discrimination since none of the arrows is small. Figure XV6.5 shows the first two discriminant axes and shows the differences between the 4 clusters.

Figure XV6.4. Axis I of the Discriminant analysis

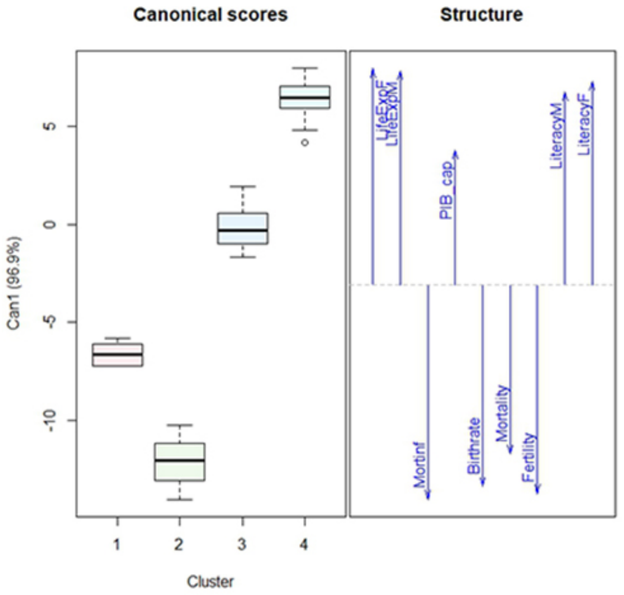
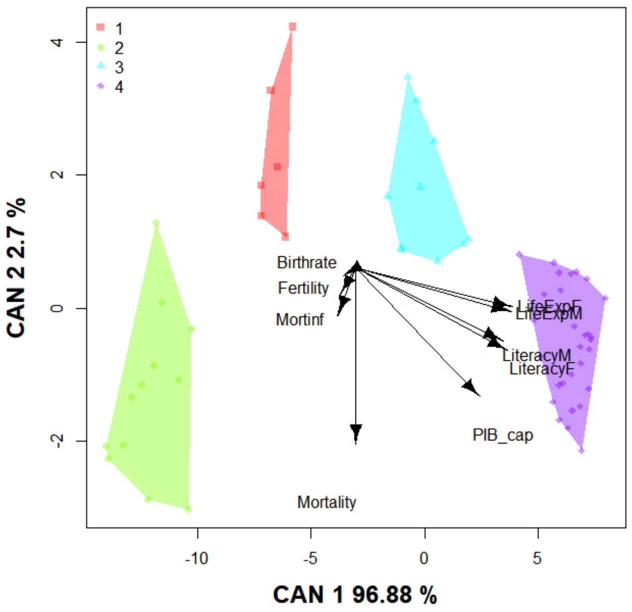


Figure XV6.5. Axes I and II of the Discriminant analysis

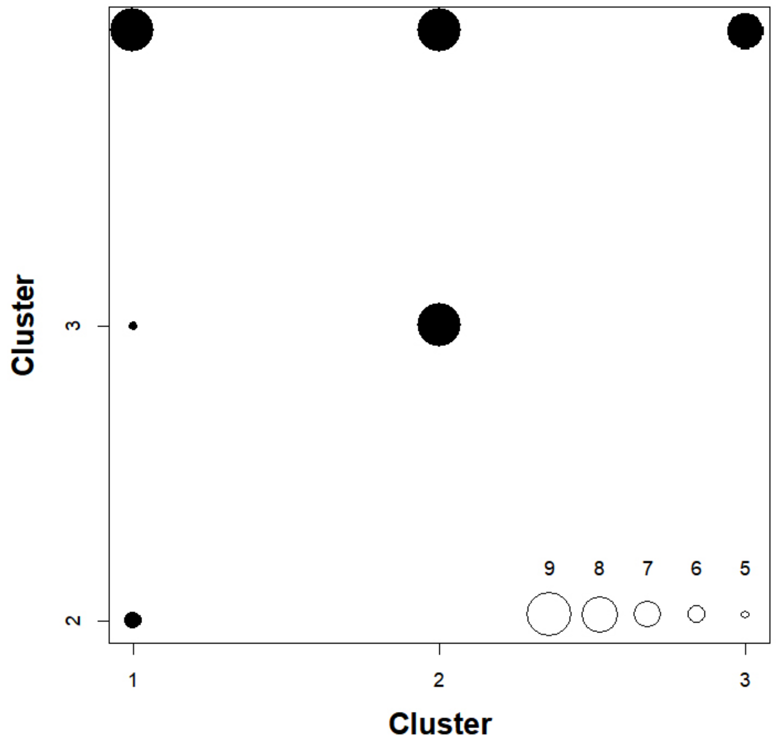


The next test to determine if the clusters are statistically different was the comparison of the variables between the clusters. The results of the U Mann-Whitney test are shown in Figure XV6.6.

For clusters to be different, there must be at least one statistically different variable when comparing each cluster with all the others. In the graph it is noted that in the comparison between all the clusters there is always a point, that is, there is always at least one variable that is different. In fact, between cluster 1 and cluster 3, the smaller number of statistically different variables was observed, a total of five variables.

Therefore, from the comparison of the variables between clusters with the U Mann-Whitney test, it is concluded that the clusters are statistically different from each other.

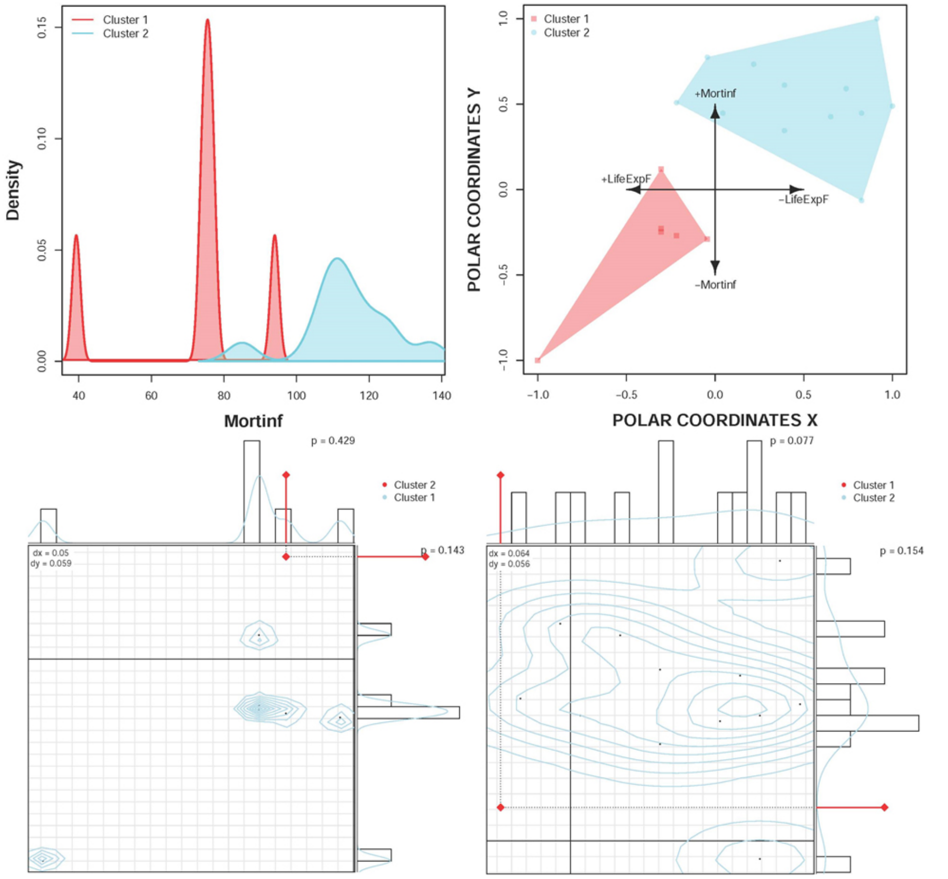
Figure XV6.6. Plot where the bubbles show the number of variables, that are statistically different (p <= 0.05) between clusters.



Finally, in a pdf, the plots obtained from applying the VARSEDIG algorithm are saved, whose objective is to compare all the clusters with each other.

Figure XV6.7 shows the example of the comparison of cluster 1 with 2. It is observed that the variables that discriminate significantly between both clusters are infant mortality and life expectancy in women (upper right panel). The Monte-Carlo test showed that the country that most resembles Cluster 2 in Cluster 1 (lower left panel) does not have significant differences in the polar coordinate axis X (p = 0.429) and in the Y axis (p = 0.143). In the country that most resembles cluster 1 to cluster 2 (bottom right panel), it is very close to the significance threshold on the polar coordinate axis X (p = 0.077) and there are no significant differences on the Y axis (p = 0,154). Therefore, it cannot be concluded that cluster 1 and 2 are different. The same process would be done to compare the rest of the clusters.

Figure XV6.7. Plots obtained from the algorithm VARSEDIG.
It is shown the comparison between the cluster 1 and 2.



Therefore, according to the Discriminant Analysis and the tests performed with the U Mann-Whitney test, the clusters are statistically different from each other, but the VARSEDIG algorithm showed that not all clusters are statistically different. However, it is very important to emphasize that the VARSEDIG algorithm considers two statistically different groups if the case that most resembles each group is statistically different using the Monte-Carlo test. The Monte-Carlo test needs a large number of cases in each group for detecting significant differences. That is, it is possible that, as it was shown in the comparison of cluster 1 with cluster 2, the cases of both groups that resemble each other are not within the point cloud of the other group, but due to the low number of cases in each group, it is not possible to determine that the difference is not due to chance.

Value

It is obtained:

1. A TXT file with the VIF (if the argument *VIF=TRUE*), the correlations between variables, the Kaiser-Meyer-Olkin (KMO) test, the Bartlett sphericity test and the results of the Principal Components analysis. The file is called by default "Output.TXT".

2. A CSV FILE with the coordinates for each case of the Principal Components analysis. The file is called by default "Cat loadings.CSV".

3. A CSV FILE with the descriptive statistics of each variable for each of the clusters obtained in the Dendrogram. The file is called by default "Descriptive statistics of clusters.CSV".

4. A CSV FILE with the original data of the variables and the cluster to which each case belongs. The file is called by default "Original data and cluster number.CSV".

5. A CSV FILE with the coordinates of the variables in the Linear Discriminant Analysis plot. The file is called by default "Var loadings-Linear.csv"

6. A CSV FILE with the coordinates of the categories in the Linear Discriminant Analysis plot. The file is called by default "Cat loadings-Linear.csv".

7. A CSV FILE with the predictions table using the cross-validation of Linear Discriminant Analysis. The file is called by default "Table cross-validation-Linear.csv".

8. A CSV FILE with the group to which each case belongs and the prediction of the Discriminant Analysis using the cross-validation of the Linear Discriminant Analysis. The file is called by default "Cases cross-validation-Linear.csv".

9. A CSV file with the predictions table using the cross-validation of the Quadratic Discriminant Analysis. The file is called by default "Table cross-validation-Quadratic.csv".

10. A CSV file with the group to which each case belongs and the prediction of the Discriminant Analysis using the cross-validation of the Quadratic Discriminant Analysis. The file is called by default "Cases cross-validation-Quadratic.csv".

11. A CSV file with the obtained probabilities of comparing all the variables among all the clusters with the U Mann-Whitney test.

12. A PDF file with the graphics obtained from the VARSEDIG algorithm.

13. A scatterplot of the Principal Components analysis.

14. A Dendrogram grouping by clusters according to the scores of the Principal Components analysis.

15. A graphic panel with a boxplot for each variable comparing the values of these variables between each of the clusters obtained in the Dendrogram.

16. A Graph of the Discriminant Analysis showing the influence of the variables on the discriminant axis I, differentiating the different clusters.

17. A graph of the Discriminant Analysis showing the scores of the discriminant axes I and II, differentiating the different clusters.

18. A bubble chart with the number of variables that are statistically different between clusters.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2018) Companion to Applied Regression. R package version 3.0-0. Available at: http://CRAN.R-project.org/package=car.

Friendly, M. (2007). HE plots for Multivariate General Linear Models. *Journal of Computational and Graphical Statistics*, 16: 421-444.

Friendly, M. & Fox, J. (2017) Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.8-0. Available at: http://CRAN.R-project.org/package=candisc.

Guisande, C., Vari, R.P., Heine, J., García-Roselló, E., González-Dacosta, J., Pérez-Schofield, B.J., González-Vilas, L. & Pelayo-Villamil, P. (2016) VARSEDIG: an algorithm for morphometric characters selection and statistical validation in morphological taxonomy. *Zootaxa*, 4162: 571-580.

Guisande, C. (2018) An Algorithm for Morphometric Characters Selection and Statistical Validation in Morphological Taxonomy. R package version 1.8. Available at: http://CRAN.R-project.org/package=VARSEDIG.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: http://CRAN.R-project.org/package=IDPmisc.

Maier, M.J. (2015) Companion Package to the Book 'R: Einführung durch angewandte Statistik. R package version 0.9.3. Available at: http://CRAN.R-project.org/package=REdaS.

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Revelle,W. (2018) Procedures for Psychological, Psychometric, and Personality Research. R package version 1.8.4. Available at: http://CRAN.R-project.org/package=psych.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2018) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-50. Available at: http://CRAN.R-project.org/package=MASS.

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17: 1-25.

Rizopoulos, D. (2018) Latent Trait Models under IRT. R package version 1.1-1. Available at: http://CRAN.R-project.org/package=ltm.

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXIII2)

XV6(data=ZXIII2, var=c("LifeExpF", "LifeExpM", "Mortinf", "PIB_cap", "Birthrate",
"Mortality", "Fertility", "LiteracyM", "LiteracyF"), cat="Continent", labels="Country",
k=4, convex=TRUE)


## End(Not run)
```

XV7 *DETECTION OF OUTLIERS IN SINGLE VARIABLES*

Description

Outliers are detected in single variables and they may also automatically deleted.

Usage

```
XV7(data, var, cat=NULL, varCode=NULL, remove=FALSE, Nan=FALSE, quant1=0.05,
quant2=0.95, file="Outliers.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Data file.
`var`	Variables to detect outliers.
`cat`	If a variable is specified, the outliers are calculated considering each group of that variable. For example, if the variable includes different species, instead of looking for the outliers based on all the data of the variable, it is only considered the data of each species separately.
`varCode`	Optionally, variables of the original matrix can be selected, which are exported in the output file with the variables in which outliers are detected. For example, this allows to choose variables which are codes of rows.
`remove`	If TRUE, the rows that have some variable with outliers are removed. If it is FALSE, the outliers are indicated with the value -9999.
`Nan`	If it is TRUE the outliers are shown as NA instead of -9999.
`quant1`	Quantile of the lower end to the elimination of outliers.
`quant2`	Quantile of the upper end to the elimination of outliers.
`file`	CSV FILE. Name of the file with the variables showing the outliers or with the variables with the outliers removed.
`na`	CSV FILE. Text that is used in the cells without data.
`dec`	CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".
`row.names`	CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

Outliers are considered those that are outside 1.5*IQR (interquartile range). IQR is the difference between the value specified in the arguments *quant2* and *quant1*. By default are 0.95 and 0.05, but for example in the boxplot are usually 0.75 and 0.25.

EXAMPLE

Outliers are detected in some morphometric variables of fish and with the argument *remove=TRUE* the rows with any outlier are removed.

Value

A CSV file is exported with the variables showing the outliers and with the variables included in the argument *varCode* if any variable was included in this argument. If the option *remove=TRUE* the rows are deleted if any variable has an outlier.

Examples

```
## Not run:

data(ZX3)

XV7(data=ZX3, var=c("M2", "M3", "M4", "M5", "M6", "M7", "M8"), cat="Species",
varCode=c("Order", "Family", "Genus", "Species"), remove=TRUE)
## End(Not run)
```

XV8 *DENDROGRAM ON A CORRESPONDENCE ANALYSIS*

Description

A Dendrogram is applied on a Correspondence analysis.

Usage

```
XV8(data, var, labels, cat=NULL, por=80, k=NULL, pthreshold=0.05,
ellipse=FALSE, convex=FALSE, dim=c(1,2), size=c(1,5), showCluster=TRUE,
method="overlap", minimum=TRUE, ResetPAR=TRUE, PAR=NULL, SCATTERPLOT=NULL,
HCLUST=NULL, CLUSTER=NULL, BOXPLOT=NULL, mfrowBOXPLOT=NULL,
LabelCat=NULL, COLOR=NULL, COLORC=NULL, COLORB=NULL,
PCH=NULL, XLIM=NULL, YLIM=NULL, XLAB=NULL, YLAB=NULL, ylabBOXPLOT=NULL,
LEGEND=NULL, MTEXT= NULL, TEXTvar=NULL, TEXTlabels=NULL, arrows=TRUE,
larrow=0.7, colArrows="black", quadratic=FALSE, file1="Output.txt",
file2="Cat loadings.csv", file3="Descriptive statistics of clusters.csv",
file4="Original data and cluster number.csv", file5="Var loadings-Linear.csv",
file6="Cat loadings-Linear.csv", file7="Table cross-validation-Linear.csv",
file8="Cases cross-validation-Linear.csv", file9="Table cross-validation-Quadratic.csv",
file10="Cases cross-validation-Quadratic.csv", file11="Plots VARSEDIG.pdf",
file12="U Mann-Whitney test.csv", na="NA", dec=",", row.names=TRUE)
```

Arguments

`data` Data file.

`var` Variables that are included in the analysis. If they are factors or characters, they are transformed into numerics and a table with the numerical value assigned to each category is shown in the results.

`labels` Variable that allows to display a label for each case.

`cat` Optionally you can specify a variable to show a grouping in the plot of the Correspondence analysis.

`por` Cut-off threshold specifying the cumulative variance percentage, to determine how many axes are selected from the Correspondence analysis. By default it is 80%, which means that the axes are selected until reaching an accumulated variance percentage of 80%.

`k` Number of clusters in which the Dendrogram is divided. If it is NULL, the algorithm select automatically the maximum number of clusters in which the Dendrogram can be divided, which are those groups that are statistically different in at least one variable according to the U Mann-Whitney test.

`pthreshold` Threshold probability of the U Mann-Whitney test.

`ellipse` If it is TRUE, the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *cat* are depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*. If it is TRUE, the ellipses of the clusters in the Discriminant analysis and in the polar coordinate plot of the VARSEDIG algorithm are also calculated.

`convex` If it is TRUE, the convex hull is calculated for each category in the Correspondence analysis, but only if some variable has been selected in the argument *cat*. If TRUE, the convex hull of the clusters is also calculated in the Discriminant analysis and in the polar coordinate plot of the VARSEDIG algorithm.

`dim` Vector with two values indicating the axes that are shown in the Correspondence analysis.

`size` Size range of bubbles. Two values: minimum and maximum size.

`showCluster` If it is TRUE, the number of each cluster is shown in the Dendrogram.

`method` Three different methods for prioritizing the variables according to their capacity for discrimination can be used in the VARSEDIG algorithm. If the method is "overlap", a density curve is obtained for each variable and the overlap of the area under the curve between the two groups of the variable *group* is estimated for all variables. Those variables with lower overlap should have better discrimination capacities and, hence, all variables are ordered from lowest to highest overlap; in other words, from the highest to lowest discrimination capacity. If the method is "Monte-Carlo", a Monte-Carlo test is performed comparing all values of group 1 with group 2, and all values of group 2 with 1. The variables are prioritized from the variable with the lowest mean of all p-values (highest discrimination capacity) to the variable with the highest mean of all p-values (lowest discrimination capacity). If the method is "logistic regression", then a binomial logistic regression is calculated and only significant variables are selected for further analyses with the regression performed by steps using the Akaike Information Criterion (AIC).

`minimum` If it is TRUE, the algorithm is designed to find a significant discrimination between both groups with the minimum possible number of significant variables. Therefore, only the variables with higher discrimination capacity are selected. If it is FALSE, the algorithm selects all significant variables, and not only those with higher discrimination capacity. This argument is only valid with the methods "Monte-Carlo" and "overlap" and it is useful in those cases that discrimination between the groups is difficult and requires to include as many as variables as possible.

`ResetPAR` If it is FALSE, the default condition of the function PAR are not placed and those defined by the user on previous graphics are maintained.

`PAR` It accesses thePAR function that allows to modify many different aspects of the graphs.

`SCATTERPLOT` It accesses the function scatterplot of the car package.

`HCLUST` You may access the function hclust of the stats package.

`CLUSTER` Access to the function that allows to modify the graphic representation of the Dendrogram.

`BOXPLOT` Allows to specify the characteristics of the boxplot.

`mfrowBOXPLOT` It allows to specify the boxplot panel. It is a vector with two numbers, for example c(2,5) which means that the boxplots are put in 2 rows and 5 columns.

`LabelCat` It allows to specify a vector with the names of the clusters in the boxplots. They must be as many as clusters.

`COLOR` It allows to modify the colours of the graphic in the Correspondence analysis, but they must be as many as different groups have the variable *cat*.

`COLORC` It allows to modify the colours of the clusters in the Dendrogram, but they must be as many as what is specified in the argument *k*.

`COLORB` It allows to modify the colours of the clusters in the boxplots, but they must be as many as what is specified in the argument *k*.

`PCH` Vector with the symbols of the Correspondence analysis plot, which must be as many as different groups have the variable *cat*. If it is NULL they are calculated automatically starting with the symbol 15.

`XLIM, YLIM` Vectors with the axes limits *X* and *Y* of the Correspondence analysis plot.

`XLAB, YLAB` Legends of the axes *X* and *Y* of the Correspondence analysis plot.

`ylabBOXPLOT` You can specify a vector with the legends of the axes *Y* of the boxplots. They should be as many as the number of variables.

`LEGEND` It allows to include or to modify a legend in the Correspondence analysis plot.

`MTEXT` It allows to add text in the margins of the Correspondence analysis plot.

`TEXTvar` It allows to modify the labels of the variables in the Correspondence analysis plot.

`TEXTlabels` It allows to modify the labels of the cases in the graph in the Correspondence analysis.

`arrows` If it is TRUE the arrows are shown in the scatterplot of the Correspondence analysis.

`larrow` It modifies the length of the arrows in the Correspondence analysis plot.

`colArrows` Colours of the arrows in the Correspondence analysis plot.

`quadratic` If TRUE, a Quadratic Discriminant Analysis is performed, in addition to the Linear Discriminant Analysis.

`file1` TXT FILE. Name of the output file with the results.

`file2` CSV FILE. Name of the output file with the coordinates of the cases of the Correspondence analysis plot.

`file3` CSV FILE. Name of the output file with the descriptive statistics of each variable for each of the clusters obtained in the Dendrogram.

`file4` CSV FILE. Name of the output file with the original data of the variables and the cluster to which each case belongs.

`file5` CSV FILE. Name of the output file with the coordinates of the variables in the Linear Discriminant Analysis plot.

`file6` CSV FILE. Name of the output file with the coordinates of the categories in the Linear Discriminant Analysis plot.

`file7` CSV FILE. Name of the output file with the prediction table using the cross-validation of the Linear Discriminant Analysis.

`file8` CSV FILE. Name of the output file with the group to which each case belongs and the prediction of the Discriminant Analysis using the cross-validation of the Linear Discriminant Analysis.

`file9` CSV FILE. Name of the output file with the predictions table using the cross-validation of the Quadratic Discriminant Analysis.

`file10` CSV FILE. Name of the output file with the group to which each case belongs and the prediction of the Discriminant Analysis using the cross-validation of the Quadratic Discriminant Analysis.

`file11` PDF File. Name of the output file with the graphics obtained from the VARSEDIG algorithm.

`file12` CSV FILE. Name of the output file with the obtained probabilities of comparing all the variables among all the clusters with the U Mann-Whitney test.

`na` CSV FILES. Text that is used in the cells without data.

`dec` CSV FILES. It defines if a comma "," or a dot "." is used as decimal separator.

`row.names` CSV FILES. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

The aim of this analysis is to determine what statistically different groups are formed by applying a Correspondence analysis.

This analysis can be useful to find different groups when you really do not know what they are. For example, find different species using qualitative variables, without really knowing how many potential species there are and to what species each individual belongs.

However, it is important to note that only different groups will be detected if the variables that have more variability give rise to different groups. It is possible that a variable does not present a great variability, but it is important for discriminating groups. This type of differentiation based on variables that do not have high variance, would not be detected in this analysis.

To detect the potential groups being formed, a Dendrogram is applied to the scores obtained from the axes that absorb a greater variance. By default, the axes that absorb 80% of the variability are chosen, but this value can be modified by the user.

Subsequently, a Discriminant Analysis is carried out to determine if the clusters that have been generated are well discriminated, that is, to determine the number of correctly identified cases in each cluster.

Next, a U Mann-Whitney test is performed to determine if there are significant differences in the variables between the clusters.

Finally, the algorithm of the VARSEDIG function is applied (see for more details Guisande et al 2016: Guisande, 2018). With this algorithm it is possible to determine if all the cases of each cluster are statistically different from the other clusters.

The idea of this function is to find the largest possible number of clusters with the highest discrimination percentage. To do this the user should perform tests, modifying the cut-off threshold by specifying the cumulative variance percentage to determine how many axes are selected from the Main Components (by default *by=80*) and the variables to be included, eliminating those that are not correlated and are not useful in the Correspondence analysis, as well as those that have little discrimination power in the Discriminant Analysis.

FUNCTIONS

The Correspondence analysis was performed with the ca function of the package ca (Greenacre & Pardo, 2006; Greenacre, 2007; Nenadic & Greenacre, 2007; Greenacre, 2013).

It was used the scatterplot function of the car package was used (Fox et al., 2018) for performing the *biplot* graph.

The arrows are depicted with the function Arrows of the package IDPmisc (Locher & Ruckstuhl, 2014).

The convex hull is estimated with the function chull of the package grDevices.

The U Mann-Whitney test is performed with the *wilcox.test* function of the base stats package.

The comparison between clusters with the VARSEDIG algorithm is done with the VARSEDIM function of the VARSEDIG package (Guisande et al., 2016: Guisande, 2018).

The Linear Discriminant Analysis was performed with the functions candisc of the candisc package (Friendly, 2007; Friendly & Fox, 2017) and lda of the MASS package (Venables & Ripley, 2002; Ripley et al., 2018).

The Quadratic Discriminant Analysis was performed with the function qda of the MASS package (Venables & Ripley, 2002; Ripley et al., 2018).

The graph with one dimension in the Discriminant analysis was performed with the function plot.cancor of the candisc package (Friendly, 2007; Friendly & Fox, 2017).

EXAMPLE

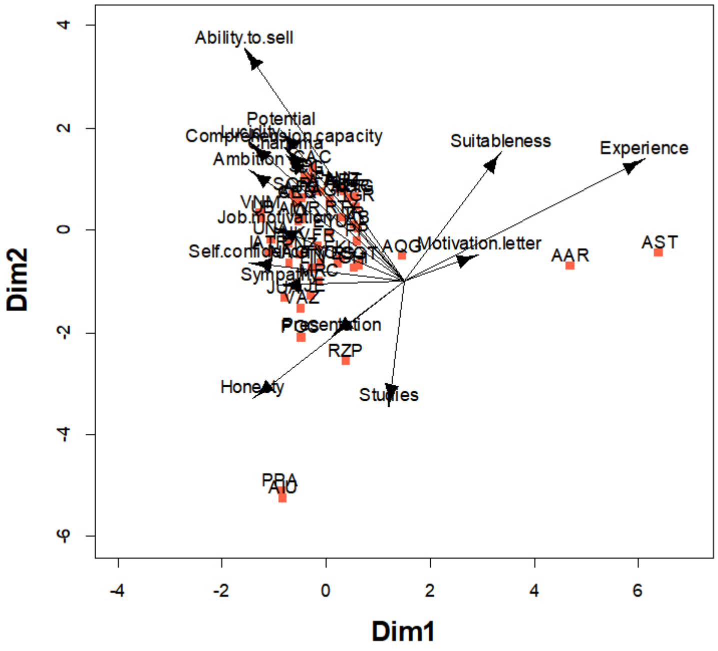
In the example, valuation criteria are used, on a scale of 0 to 10, of 48 candidates who apply for a job. The objective is to see which groups are formed from the Correspondence analysis.

Figures XV8.1 shows that in the right end of the first axis the variables are experience, the suitability for the job and letter of motivation.

In the left end of the first axis motivation and confidence are associated. The second axis is associated with the business sense in the upper end, while at the lower end are honesty and level of studies.

The eigenvalues show the variance percentage of the axes: 32.9% first axis, 25.7% the second, 11.3% the third, etc.

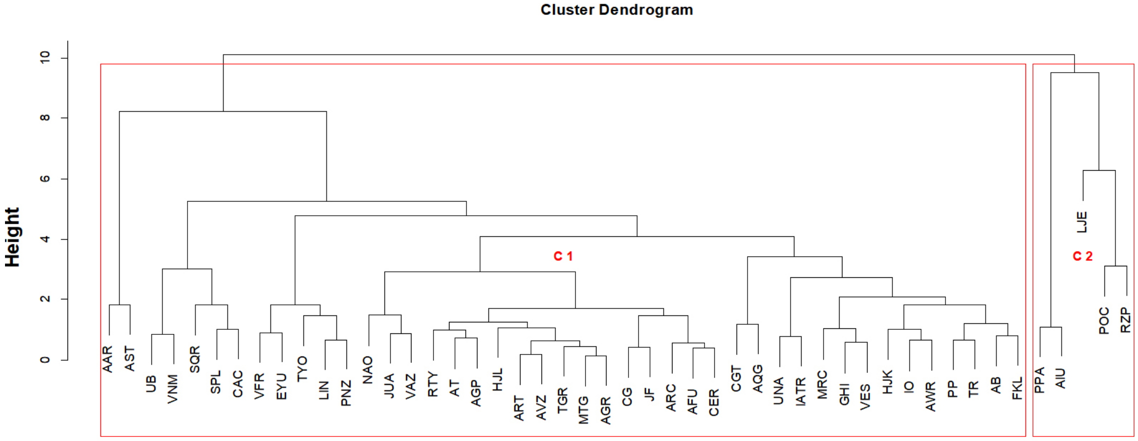
Figure XIII.1. Correspondence analysis showing the observed variability in the qualities of candidates for a job.



Inertia (*Inertia*) indicates the relative weight or the different importance of each variable. The highest values are for the experience (0.0271), level of studies (0.0168), honesty (0.0155), motivation letter (0.015), adequacy (0.0149), commercial sense (0.012, *Ability.to.sell*), etc.

In the example, the option *kmethod= automatically"* was chosen and it is observed in Figure XV8.2 that there are only 2 statistically significant clusters with at least one variable different according to the Mann-Whitney U test.

Figure XV8.2. Dendrogram with the scores of the axes selected from the Correspondence analysis.



In Figure XV8.3 the differences between clusters for each of the variables are observed. Both clusters differ significantly in almost all variables, specifically in 13 of them (Figure XV8.4). There are only no differences in honesty and in the level of studies.

Figure XV8.3. Boxplot obtained for each of the variables
with the averaged values for each cluster.

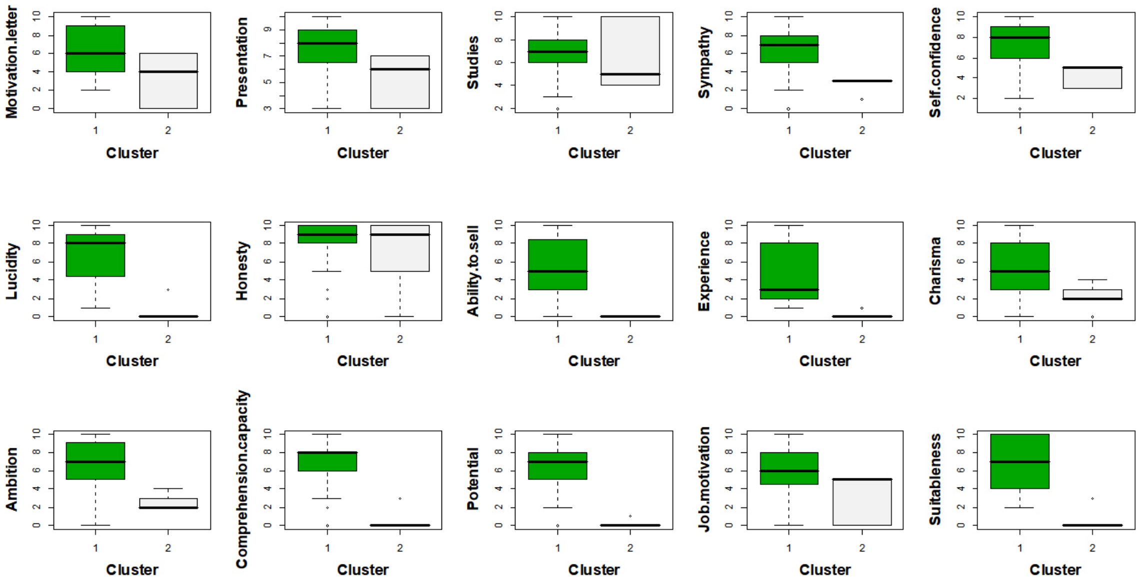
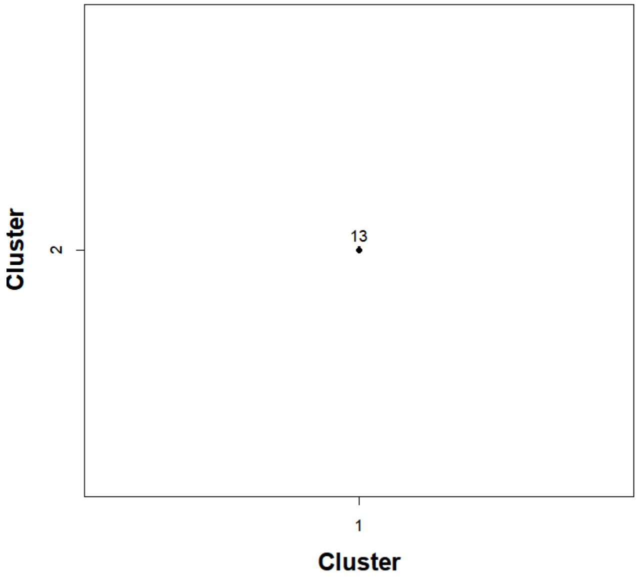
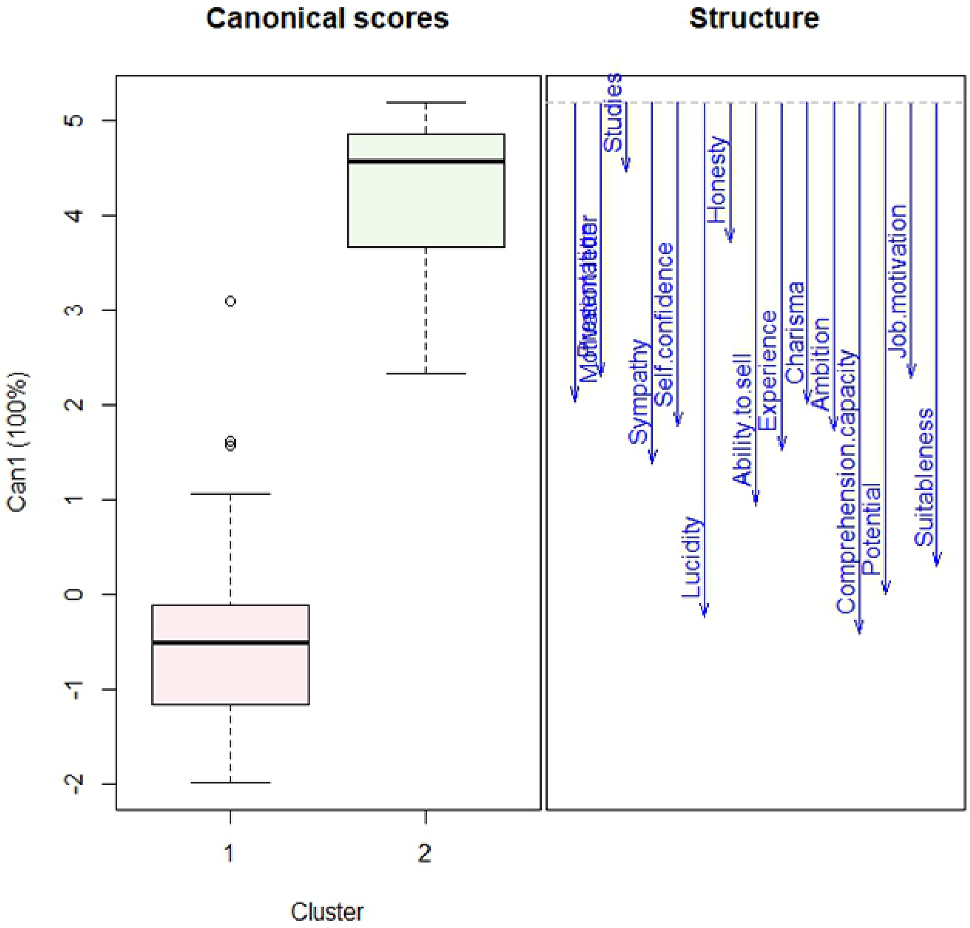


Figure XV8.4. Plot where the bubbles represent the number of variables.
that are statistically different (p <= 0.05) between clusters.



The Discriminant analysis shows that it is possible to correctly discriminate 87.5% of cases by cross-validation with the Linear method. All the variables seem to be important for discrimination, with the exception of the level of studies and honesty, in which the arrows are small (Figure XV8.5).

Figure XV8.5. Axis I of the Discriminant analysis.

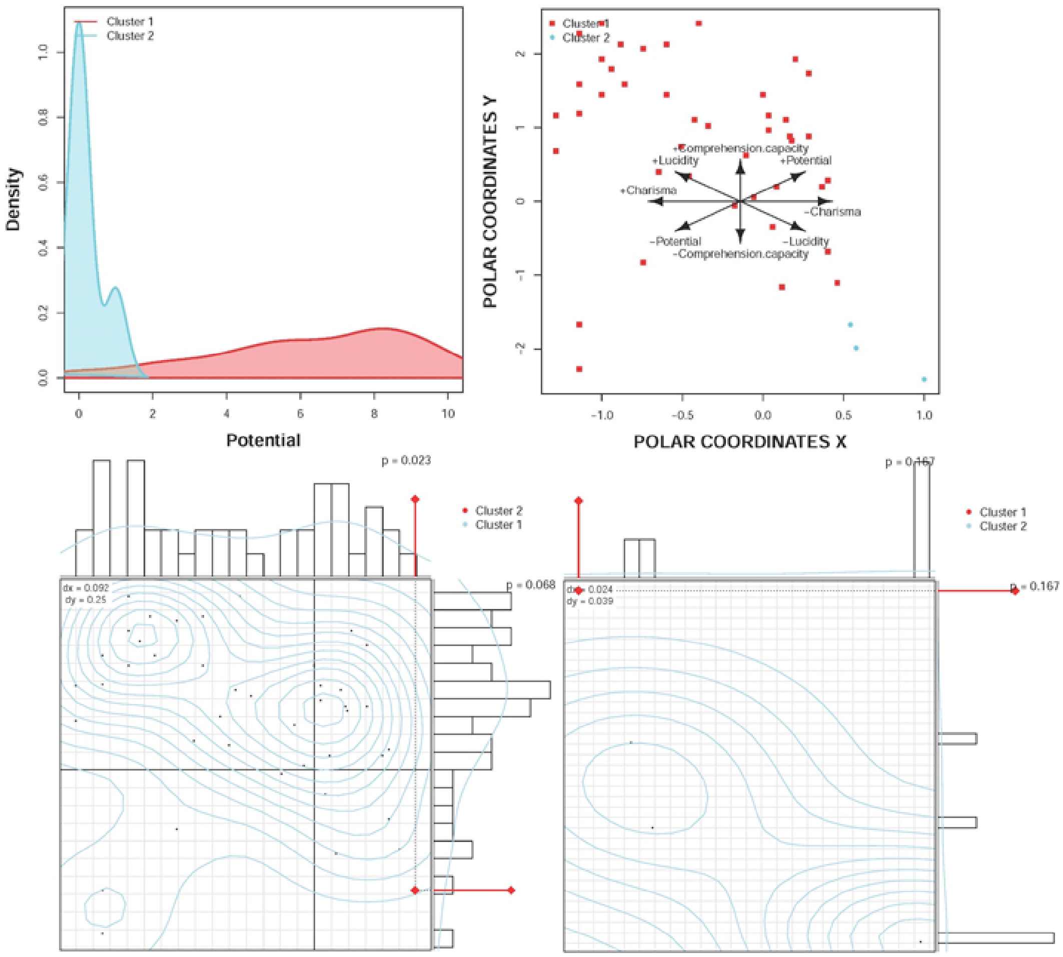


Finally, in a pdf, the graphics obtained from applying the VARSEDIG algorithm are saved. The aim of this algorithm is to compare all the clusters with each other.

Figure XV8.6 shows the example of the comparison of cluster 1 with 2. It is observed that the variables that discriminate significantly between both clusters are the potential, comprehensibility, lucidity and charisma (upper right panel).

The Monte-Carlo test showed that the candidate that most resembles cluster 2 in cluster 1 (lower left panel) there are significant differences on the X axis (p = 0.023) and almost on the Y axis (p = 0.068). However, the candidate that most resembles cluster 1 to cluster 2 does not have differences on the X axis (p = 0.167) nor on the Y axis (p = 0.167). Therefore, it cannot be concluded that cluster 1 and 2 are different.

Figure XV8.6. Plots obtained from VARSEDIG algorithm. It is shown the comparison of the cluster 1 with the cluster 2.



Therefore, according to the tests carried out with the U Mann-Whitney test, the clusters are statistically different from each other. In the Discriminant, a complete separation of both groups was not achieved. In the algorithm VARSEDIG showed that all the candidates from cluster 2 are different from cluster 1, but not all from cluster 1 are different from cluster 2. However, it is very important to note that the VARSEDIG algorithm considers two statistically different groups if the case that the more it resembles each group the other, it is statistically different using the Monte-Carlo test. The Monte-Carlo test needs a large number of cases in each group for detecting significant differences. In the case of group 2 the number of candidates is very small and, therefore, it is not possible to determine that the differences are not by chance.

Value

It is obtained:

1. A TXT file with the results of the Correspondence analysis. The file is called by default "Output.TXT".

2. A CSV FILE with the coordinates for each case of the Correspondence analysis. The file is called by default "Cat loadings.CSV".

3. A CSV FILE with the descriptive statistics of each variable for each of the clusters obtained in the Dendrogram. The file is called by default "Descriptive statistics of clusters.CSV".

4. A CSV FILE with the original data of the variables and the cluster to which each case belongs. The file is called by default "Original data and cluster number.CSV".

5. A CSV FILE with the coordinates of the variables in the Linear Discriminant Analysis plot. The file is called by default "Var loadings-Linear.csv"

6. A CSV FILE with the coordinates of the categories in the Linear Discriminant Analysis plot. The file is called by default "Cat loadings-Linear.csv".

7. A CSV FILE with the predictions table using the cross-validation of Linear Discriminant Analysis. The file is called by default "Table cross-validation-Linear.csv".

8. A CSV FILE with the group to which each case belongs and the prediction of the Discriminant Analysis using the cross-validation of the Linear Discriminant Analysis. The file is called by default "Cases cross-validation-Linear.csv".

9. A CSV file with the predictions table using the cross-validation of the Quadratic Discriminant Analysis. The file is called by default "Table cross-validation-Quadratic.csv".

10. A CSV file with the group to which each case belongs and the prediction of the Discriminant Analysis using the cross-validation of the Quadratic Discriminant Analysis. The file is called by default "Cases cross-validation-Quadratic.csv".

11. A CSV file with the obtained probabilities of comparing all the variables among all the clusters with the U Mann-Whitney test.

12. A PDF file with the graphics obtained from the VARSEDIG algorithm.

13. A scatterplot of the Correspondence analysis.

14. A Dendrogram grouping by clusters according to the scores of the Correspondence analysis.

15. A graphic panel with a boxplot for each variable comparing the values of these variables between each of the clusters obtained in the Dendrogram.

16. A Graph of the Discriminant Analysis showing the influence of the variables on the discriminant axis I, differentiating the different clusters.

17. A graph of the Discriminant Analysis showing the scores of the discriminant axes I and II, differentiating the different clusters.

18. A bubble chart with the number of variables that are statistically different between clusters.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2018) Companion to Applied Regression. R package version 3.0-0. Available at: http://CRAN.R-project.org/package=car.

Friendly, M. (2007). HE plots for Multivariate General Linear Models. *Journal of Computational and Graphical Statistics*, 16: 421-444.

Friendly, M. & Fox, J. (2017) Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.8-0. Available at: http://CRAN.R-project.org/package=candisc.

Greenacre, M. (2013). Simple, Multiple and Joint Correspondence Analysis. R package version 0.53. Available at: http://CRAN.R-project.org/package=ca.

Greenacre, M. (2007) *Correspondence Analysis in Practice*. Second Edition. London: Chapman & Hall / CRC.

Greenacre, M.J. & Pardo, R. (2006) Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods and Research*, 35: 193-218.

Guisande, C., Vari, R.P., Heine, J., García-Roselló, E., González-Dacosta, J., Pérez-Schofield, B.J., González-Vilas, L. & Pelayo-Villamil, P. (2016) VARSEDIG: an algorithm for morphometric characters selection and statistical validation in morphological taxonomy. *Zootaxa*, 4162: 571-580.

Guisande, C. (2018) An Algorithm for Morphometric Characters Selection and Statistical Validation in Morphological Taxonomy. R package version 1.8. Available at: http://CRAN.R-project.org/package=VARSEDIG.

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: http://CRAN.R-project.org/package=IDPmisc.

Nenadic, O. & Greenacre, M. (2007) Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20: 1-13.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. & Firth, D. (2018) Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-50. Available at: http://CRAN.R-project.org/package=MASS.

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*. Springer, fourth edition, New York. http://www.stats.ox.ac.uk/pub/MASS4.

Examples

```
## Not run:

data(ZXIII1)

XV8(data=ZXIII1, var=c("Motivation.letter", "Presentation", "Studies",
"Sympathy", "Self.confidence", "Lucidity", "Honesty", "Ability.to.sell",
"Experience", "Charisma", "Ambition", "Comprehension.capacity",
"Potential", "Job.motivation", "Suitableness"), labels="Candidate",
XLIM=c(-4,7), YLIM=c(-6,4))

## End(Not run)
```

XV9 *DETECTION OF OUTLIERS IN A COMBINATION OF VARIABLES*

Description

Outliers are detected in a combination of variables and they may also automatically deleted.

Usage

```
XV9(data, var, cat=NULL, varCode=NULL, remove=FALSE, Nan=FALSE,
quant=0.95, file="Outliers.csv", na="NA", dec=",", row.names=FALSE)
```

Arguments

`data`	Data file.
`var`	Variables to be combined for detecting outliers.
`cat`	If a variable is specified, the outliers are calculated considering each group of the combination of variables. For example, if the variable includes different species, instead of looking for the outliers based on all the data of the variable, it is only considered the data of each species separately.
`varCode`	Optionally, variables of the original matrix can be selected, which are exported in the output file with the variables in which outliers are detected. For example, this allows to choose variables which are codes of rows.
`remove`	If it is TRUE, the rows that have some variable with outliers are deleted in the output file, but the outliers are still shown in the graph. If it is FALSE, the outliers are indicated with the value -9999.
`Nan`	If it is TRUE the outliers are shown as NA instead of -9999.
`quant`	Quantile of the upper end to the elimination of outliers.
`file`	CSV FILE. Name of the file with the variables showing the outliers or with the variables with the outliers removed.
`na`	CSV FILE. Text that is used in the cells without data.
`dec`	CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".
`row.names`	CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

In some statistical analyses such as MANOVA and Principal Component analysis, we work with a combination of variables, instead of with the variables independently. It is possible that some variables do not have outliers when they are analysed independently, but the combination of them may result in atypical values. Therefore, a statistical method is needed to detect outliers when combining variables and the Mahalanobis distance is frequently used for that purpose (Mahalanobis, 1936).

The Mahalanobis distance is a measure of the distance between a point P and a distribution D. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D. This distance is zero if P is at the mean of D, and grows as P moves away from the mean along each principal component axis, the Mahalanobis distance measures the number of standard deviations from P to the mean of D. If each of these axes is re-scaled to have unit variance, then the Mahalanobis distance corresponds to standard Euclidean distance in the transformed space. The Mahalanobis distance is thus unitless, scale-invariant, and takes into account the correlations of the data set. The equation is the following:

$$D^2 = (x - \mu)^{'} C^{-1} (x - \mu)$$

where C^{-1} is the inverse covariance matrix and μ is the center.

Once the distance is calculated we have the problem of how to detect the outliers. For relatively large samples, the estimated Mahalanobis distances have a distribution that approximates a chi-square. This result can be used to evaluate (subjectively) whether a data point can be an outlier. For this, a Q-Q plot is usually used to represent the Mahalanobis distances of the sample. The basic idea is the same as in a Normality distribution graph.

For multivariate data, the ordered Mahalanobis distances versus quantiles estimated for a sample of size *n* are plotted from a chi-square distribution with *p* degrees of freedom (number of variables that are combined). This should resemble a straight line for data from a normal multi-variable distribution. The outliers will be shown as points on the upper right side of the graph for which the Mahalanobis distance is markedly larger than the chi-square quantile value.

The quantiles are determined as follows:

1. The *ith* estimated quantile is determined as the chi-square value (with df=p) for which the cumulative probability is $(i - 0.5)/n$.

2. To determine the full set of estimated chi-square quantiles, this is done for value of *i* from 1 to *n*.

The problem is that the method is subjective, that is, it is up to the user to decide, based on the graph obtained, which points are eliminated. To eliminate this subjectivity, in this function an algorithm is introduced, which using the described criterion to detect outliers in independent variables defined in the function XV7, those Mahalanobis distances that are outside the upper quantile (by default 0.95) are considered outliers.

The values considered to be atypical are highlighted in red in the Q-Q plot, so that the user can decide if the selection has been correct.

FUNCTIONS

The Mahalanobis distance is calculated with the mahalanobis function of the base stats package.

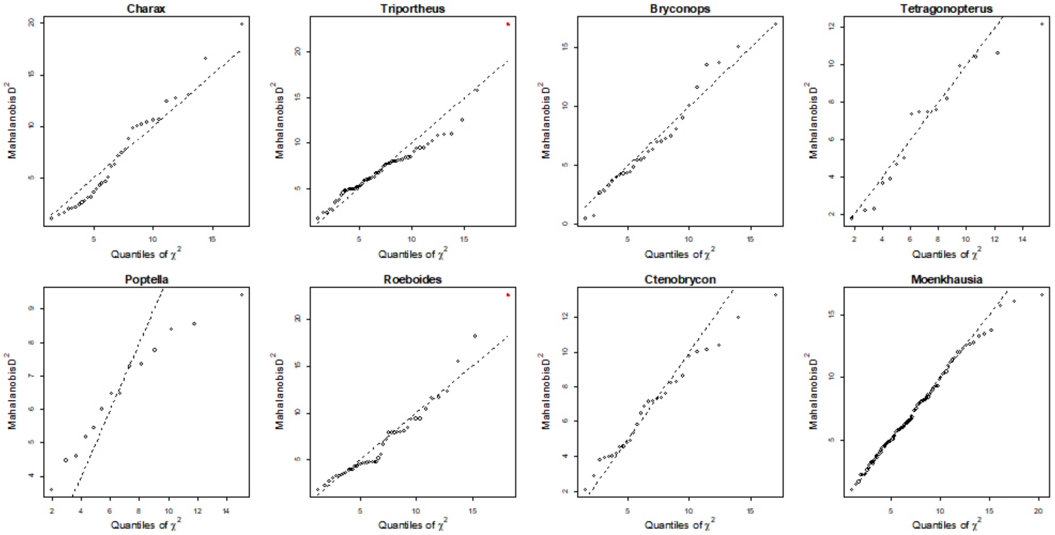
EXAMPLE

Outliers are detected in the combination of morphometric variables of fish. The outliers are shown with the -9999 value in the obtained file.

As in the script it was selected to perform the calculations by genus, *cat="Genus"*, a Q-Q plot is depicted for each genus.

In the Q_Q plots shown below, in which the Mahalanobis distance is related to the quantiles estimated from a chi-square distribution with 7 degrees of freedom (since there are 7 variables to be combined), only ouliers are detected in the genera *Triporheus* and *Roeboides*, with the default quantile of 0.95.

It is interesting to note that in the case of the genus *Poptella*, since the data number is not very large, the points do not fit the straight line of the chi-square distribution of the Q-Q plot.



Value

It is exported:

1. A CSV file with the independent variables and the Malahanobis distance resulting from the combination of the variables, where the outliers are shown. The variables included in the argument *varCode* are also exported if any variable was included in this argument. If the option *remove=TRUE* the rows are eliminated if the distance of Malahanobis is an outlier.

2. A Q-Q plot with the ordered Mahalanobis distances versus estimated quantiles (percentiles) for a sample of size *n* from a chi-squared distribution with *p* degrees of freedom (number of variables). The outliers are highlighted in red on the plot. If you chose to perform the estimation by categories, a Q-Q plot is shown for each category.

References

Mahalanobis, P.C. (1936) On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2: 49-55.

Examples

```
## Not run:
data(ZX3)

XV9(data=ZX3, var=c("M2", "M3", "M4", "M5", "M6", "M7", "M8"), cat="Genus",
varCode=c("Order", "Family", "Genus", "Species"))

## End(Not run)
```

XVI1 *Multivariate analysis of variance (MANOVA)*

Description

An analysis of variance is applied considering several dependent quantitative variables.

Usage

```
XVI1(data, var, Factor, SS="III", trans=NULL, pthreshold=0.05, psig=FALSE,
ellipse=FALSE, convex=TRUE, PCH=rep(16,100), order=TRUE, cor=TRUE,
file1="Output.txt", file2="Multinormality.csv", file3="Post hoc pvalues.csv",
na="NA", dec=",", row.names=TRUE)
```

Arguments

`data` Data file.

`var` Dependent variables.

`Factor` Variables with the factors.

`SS` When there are several factors, the decomposition of the sum of squares in parts attributed to each one of the factors is not unique. Four types of sums of squares are commonly used: type I for nested models, type II for balanced models, type III for unbalanced models (also balanced), and type IV for models with empty boxes. The type III is suitable for most applications, so it will be the one used by default. This description is very simplified; interested readers should consult supplementary bibliography for a correct use of the ANOVA models.

`trans` Type of transformation that is applied to the data:

1. NULL (untransformed)
2. 1/x2
3. 1/x
4. LN
5. LOG
6. SQR (square root)
7. x2
8. x3
9. EXP (exponential)
10. ASN (arcsine)

`pthreshold` Threshold probability that is used to group the categories of each factor, based on the probabilities obtained in the post hoc analysis.

`psig` If TRUE, in the post hoc inter-group comparison file, only those that are significant with a probability value equal to or less than the cut-off value specified in the argument *pthreshold* are shown.

`ellipse` If it is TRUE, the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the factors are depicted.

`convex` If it is TRUE, the convex hull is depicted for each category.

`PCH` Symbols of each category of the factors.

`order` If it is TRUE, the first variable of the polar coordinates is the one that best discriminates the categories. The objective is to show the greatest possible separation between the categories in the plots. If it is FALSE, the first variable is the one specified by the user in the argument *var*.

`cor` If TRUE, the variables are ordered according to the correlation between them when estimating the polar coordinates. Therefore, the next variable to another variable is the one with the highest positive correlation. The objective is to show the greatest possible separation between the categories in the plots. If it is FALSE and the argument *order=FALSE*, the variables are organized as they appear in the argument *var*.

`file1` TXT FILE. Name of the output file with the results.

`file2` CSV FILE. Name of the file with the probabilities obtained from the multinormality analysis.

`file3` CSV FILE. Name of the file with the probabilities obtained from the post hoc comparison between groups for each factor.

`na` CSV FILE. Text that is used in the cells without data.

`dec` CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".

`row.names` CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

Multivariate analysis of variance (MANOVA)

The multivariate analysis of variance (MANOVA) is an extension of the univariate analysis of variance (ANOVA). In an ANOVA, we examine the statistical differences in a continuous dependent variable between the means of the categories of one or several factors. MANOVA extends this analysis taking into account multiple continuous dependent variables, and groups them into a weighted linear combination or compound variable. The MANOVA will compare whether the newly created combination differs or not according to the different groups or levels of each factor. In this way, MANOVA essentially tests whether the independent grouping variable (factor) explains a statistically significant amount of variance in the dependent variable created as a linear combination of all independent variables.

To perform a MANOVA it is necessary to fulfill the following assumptions:

· The dependent variables are continuous, random and independent, that is, there should be no relationship between the groups or within the groups of each factor. For example, it is not valid for the data of a group to derive from the data of another group (related samples) or to repeat data between groups.

· There are no univariate or multivariate outliers. The function XV7 can be used to detect the outliers within each dependent variable, and for the multivariate outliers, that is, those that result from the combination of variables, it can be used the function XV9, both of StatR. The multivariate outliers are cases that have an unusual combination of scores in the dependent variables and, therefore, the combination that results from the dependent variables gives rise to an outlier.

· There must not be collinearity among the quantitative dependent variables. A value greater than around 10-12 of VIF would mean that there is collinearity.

· The relationship between the quantitative dependent variables is linear.

· It assumes multivariate normality, which means that the residuals are normal for each dependent variable, in each of the groups or categories of factors.

· Equality of the variances-covariances matrices is assumed in the residuals of the groups of the factors considered.

An original contribution of this function is that post hoc tests are carried out, which allows grouping the categories that are not statistically different and, therefore, showing how many real groups are different within each factor. Only post hoc tests are shown if there are more than two categories in the factor and if there are differences between some of the categories of the factor.

The algorithm consists of performing the MANOVA within each factor, but instead of comparing all the groups, each group is compared with the rest. The probabilities obtained are recorded in the file that is called by default "Post hoc pvalues.CSV". Subsequently, a matrix is generated where the rows are the categories and the columns are also categories. In each row it is indicated with a 1 the categories that are statistically different and with 0 if they are not. Therefore, each row shows with 1 and 0, what categories are or are not statistically different from the category that is considered in that row. This matrix allows us to know which categories are more similar to each other, depending on whether they have more significant categories in common.

This matrix is used to perform successive dendrograms, where different number of clusters are used. A MANOVA is applied to the clusters that are being formed and the algorithm stops when there are no longer any differences between any of the clusters. Therefore, the algorithm allows obtaining the largest possible number of different clusters within each factor, grouping in each cluster the categories that are not statistically different.

FUNCTIONS

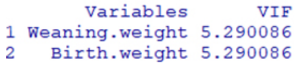
The lillie.test function of the nortest package (Gross, 2015) is used to perform the Kolmogorov-Smirnov Normality test with the Lilliefors correction. The shapiro.test function is used to perform the Shapiro-Wilk test, the linear model is performed with the lm function, both of the base stats package. The MANOVA type II-III is performed with the Manova function of the car package (Fox et al., 2018a). The MANOVA type I is done with the function manova of the base stats package. To determine the homogeneity of variances between the categories of the factors, the Box's M test was applied, for which the function boxM of the heplots package (Fox et al., 2018b) was used. The function vif of the usdm package (Naimi, 2013; Naimi et al., 2014) was used for the VIF estimation.

EXAMPLE

The data are birth weight and weaning weight of calves. Individuals at weaning are different from those whose birth weight was measured. Therefore, the dependent variables are not related. The objective is to determine if there are significant differences in both weights depending on the sex or the father. The ultimate goal is to select parents that have calves with greater weight.

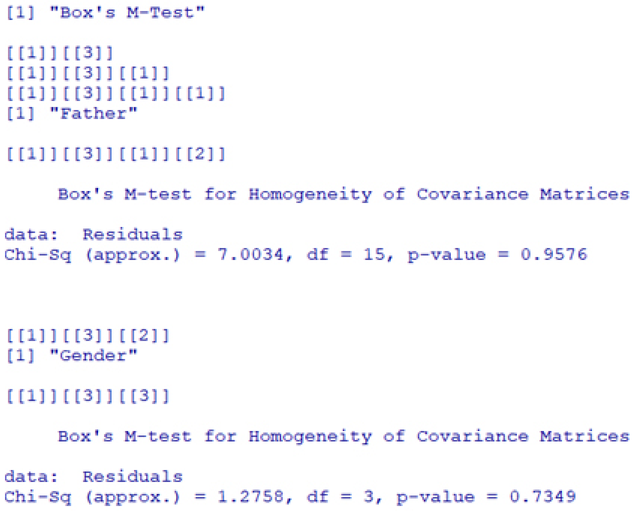
If an ANOVA is performed on each of the dependent variables, birth weight and weaning weight, there are no significant differences in birth weight between parents (p = 0.726) and sexes (p = 0.201). There are also no significant differences in weaning weight between parents (p = 0.109) and sexes (p = 0.431). However, the MANOVA shows something different.

In the TXT file of results, the first thing that is shown are the values of VIF, which are being small, so there is no collinearity among the dependent variables, which is one of the assumptions of MANOVA.



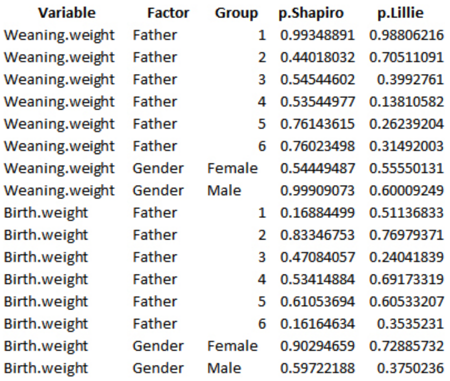
```
        Variables      VIF
1 Weaning.weight 5.290086
2   Birth.weight 5.290086
```

The requirement of equality of variances between groups of factors is also fulfilled. The Box's M test shows that there is equality in the variances-covariances matrices in the residuals of the groups of the factors considered, both in the case of gender (p = 0.734) and among the parents (p = 0.957).



```
[1] "Box's M-Test"

[[1]][[3]]
[[1]][[3]][[1]]
[[1]][[3]][[1]][[1]]
[1] "Father"

[[1]][[3]][[1]][[2]]

    Box's M-test for Homogeneity of Covariance Matrices

data:  Residuals
Chi-Sq (approx.) = 7.0034, df = 15, p-value = 0.9576


[[1]][[3]][[2]]
[1] "Gender"

[[1]][[3]][[3]]

    Box's M-test for Homogeneity of Covariance Matrices

data:  Residuals
Chi-Sq (approx.) = 1.2758, df = 3, p-value = 0.7349
```

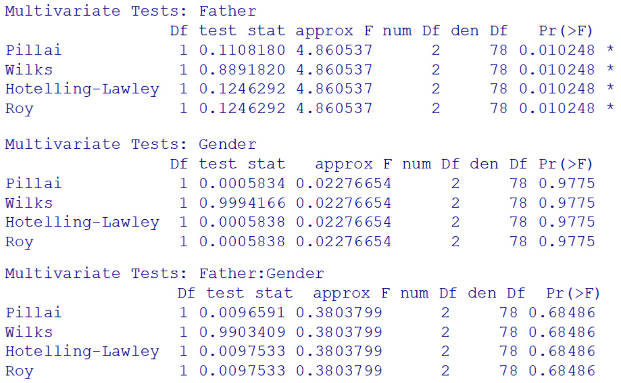
In the file that is called by default "Multinormality.csv", it is observed that the multinormality is also fulfilled. The samples are normal in the two variables for all groups of the two factors.



Variable	Factor	Group	p.Shapiro	p.Lillie
Weaning.weight	Father	1	0.99348891	0.98806216
Weaning.weight	Father	2	0.44018032	0.70511091
Weaning.weight	Father	3	0.54544602	0.3992761
Weaning.weight	Father	4	0.53544977	0.13810582
Weaning.weight	Father	5	0.76143615	0.26239204
Weaning.weight	Father	6	0.76023498	0.31492003
Weaning.weight	Gender	Female	0.54449487	0.55550131
Weaning.weight	Gender	Male	0.99909073	0.60009249
Birth.weight	Father	1	0.16884499	0.51136833
Birth.weight	Father	2	0.83346753	0.76979371
Birth.weight	Father	3	0.47084057	0.24041839
Birth.weight	Father	4	0.53414884	0.69173319
Birth.weight	Father	5	0.61053694	0.60533207
Birth.weight	Father	6	0.16164634	0.3535231
Birth.weight	Gender	Female	0.90294659	0.72885732
Birth.weight	Gender	Male	0.59722188	0.3750236

The following results are the MANOVA analysis. There are four tests that are used more frequently: Pillai, Wilks, Hotelling-Lawley and Roy. None of these tests can be clearly identified as the best test to be used in all situations (Lee, 1971, Pillai and Jayachandran, 1967). The comparative effectiveness of each of these tests changes in relation to the specific characteristics of the data. However, taking into account a wide spectrum of real data conditions, the Pillai, Wilks and Hotelling-Lawley tests generally work in a similar way, particularly when the assumptions are met (Johnson & Wichern, 2002; Finch & French, 2013).

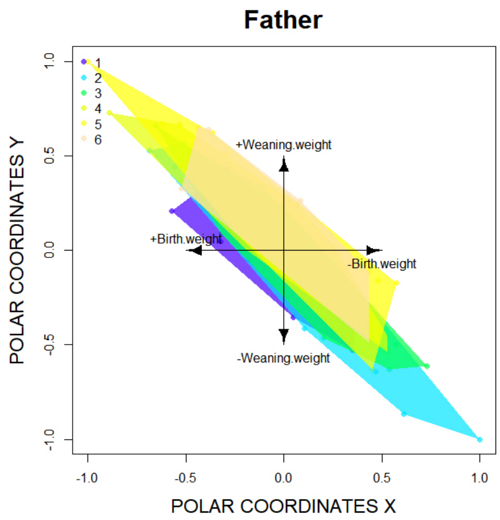
The following table shows that in all cases the test values coincide. In the case of the father factor, there are significant differences, which means that the weight of the calves at birth and at weaning varies depending on the father (p = 0.010). However, there are no significant differences between sexes (p = 0.977) nor in the parent-gender interaction (p = 0.684). Non-interaction means that calves born from different parents do not have a different weight depending on whether they are male or female. In other words, significant interaction would mean that in some parents, male calves weigh more than females, and vice versa in other parents.



```
Multivariate Tests: Father
                 Df test stat approx F num Df den Df   Pr(>F)
Pillai            1 0.1108180 4.860537      2     78 0.010248 *
Wilks             1 0.8891820 4.860537      2     78 0.010248 *
Hotelling-Lawley  1 0.1246292 4.860537      2     78 0.010248 *
Roy               1 0.1246292 4.860537      2     78 0.010248 *

Multivariate Tests: Gender
                 Df test stat   approx F num Df den Df Pr(>F)
Pillai            1 0.0005834 0.02276654      2     78 0.9775
Wilks             1 0.9994166 0.02276654      2     78 0.9775
Hotelling-Lawley  1 0.0005838 0.02276654      2     78 0.9775
Roy               1 0.0005838 0.02276654      2     78 0.9775

Multivariate Tests: Father:Gender
                  Df test stat  approx F num Df den Df  Pr(>F)
Pillai            1 0.0096591 0.3803799      2     78 0.68486
Wilks             1 0.9903409 0.3803799      2     78 0.68486
Hotelling-Lawley  1 0.0097533 0.3803799      2     78 0.68486
Roy               1 0.0097533 0.3803799      2     78 0.68486
```

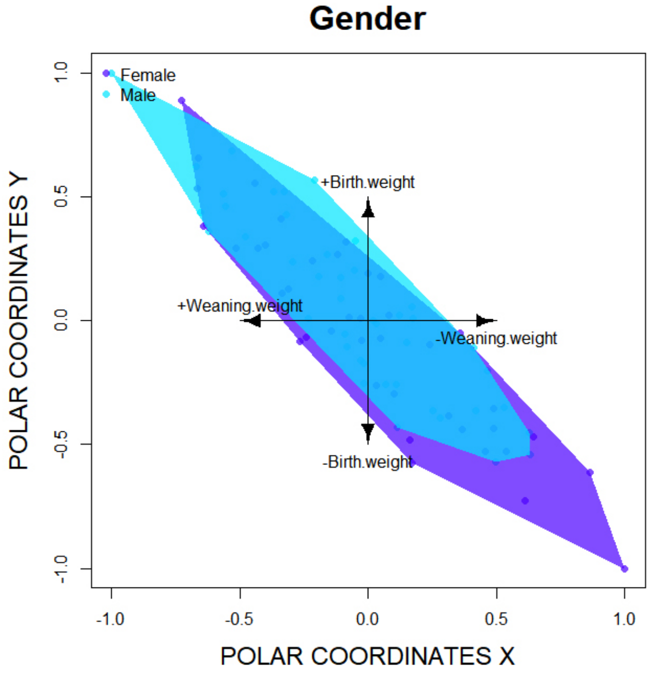
Figure XVI1.1 shows the groups of parents in the graph of polar coordinates where it is observed that there are different groups, which corroborates the differences between parents that showed the MANOVA in the weights at birth and at weaning.

Figure XVI1.1. Polar coordinates of the different parents depending on the two dependent variables analysed.

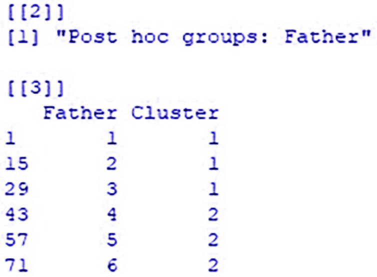


However, in the case of gender (Figure XVI1.2) two overlapping groups are observed, which also allows visualizing what is shown in the MANOVA, that there are no differences between sexes in the weight of the calves at birth and at weaning.

Figure XVI1.2. Polar coordinates of the factor gender depending on the two dependent variables analysed.



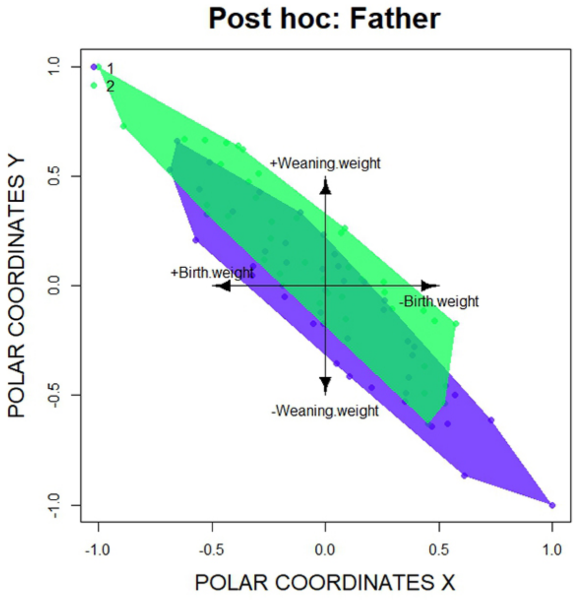
In the function, an algorithm has been implemented that performs post hoc tests, which are detailed in the *details* section. Only post hoc tests are shown if there are more than two categories in the factor and if there are differences between some of the factor categories. Therefore, in this example, there is no information about post hoc tests of the gender factor, since there are only two categories. In the case of the parent factor, two groups are generated. Group 1 includes parents 1, 2 and 3, and group 2 includes parents 4, 5 and 6, as can be seen in the table shown below and appears in the results file. Therefore, there are two significantly different groups of parents (Figure XVI1.3).



```
[[2]]
[1] "Post hoc groups: Father"

[[3]]
   Father Cluster
1       1       1
15      2       1
29      3       1
43      4       2
57      5       2
71      6       2
```

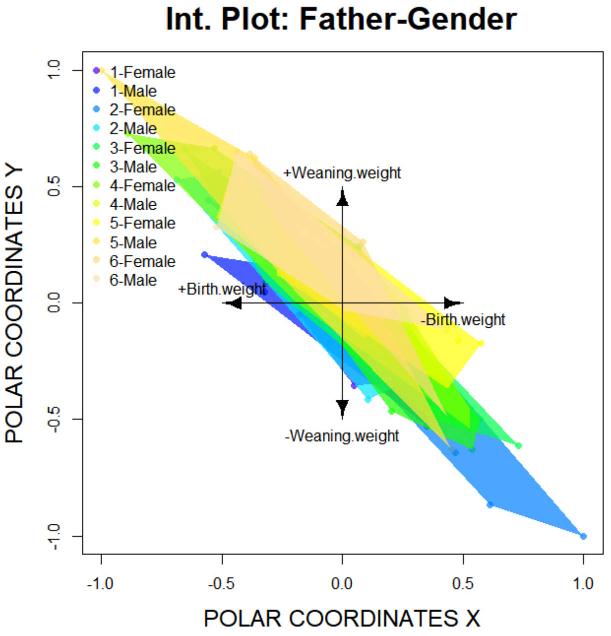
Figure XVI1.3 shows the post hoc chart for the parent factor. To obtain the graph with the ellipses, it is necessary to run the script by adding the arguments *convex = FALSE* and *ellipse = TRUE*. The central point of the ellipses of each group allows to better differentiate that group 2 has a higher average weight, both at birth and at weaning.

Figure XVI1.3. Polar coordinates of the resulting groups of the parent factor, which show significant differences in the post hoc analysis.



The last graph is that of interaction between both factors (Figure XVI1.4). It is more difficult to visualize, but it can be seen that, for the same father, for example father 6, the values are very similar in males and females. This is similar for the rest of the parents, which makes it possible to visualize that there is no significant interaction between the factors.

Figure XVI1.4. Polar coordinates of the interaction between both factors



All these polar coordinate graphs can be modified to suit the user using the original data and the F46 function to perform polar coordinates available in the PlotsR package.

Value

It is exported:

1. A TXT file with the results of the VIF, the Box' M test and the results of the MANOVA.

2. A CSV file with the probabilities obtained from the multinormality analysis.

3. A CSV file with the probabilities obtained from the post hoc comparison between the groups of each factor.

4. Polar coordinate plots showing the separation between categories for each factor, the groups resulting from the post hoc analysis and the interactions between factors.

References

Finch, H & French, B. (2013) A Monte Carlo comparison of robust MANOVA. *Test Statistics Journal of Modern Applied Statistical Methods*, 12: 35-81.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2018a) Companion to Applied Regression. R package version 3.0-0. Available at: http://CRAN.R-project.org/package=car.

Fox, J, Friendly, M. & Monette, G. (2108b) Visualizing Hypothesis Tests in Multivariate Linear Models. R package version 1.3.5. Available at: http://CRAN.R-project.org/package=heplots.

Gross, J. (2015) Tests for Normality. R package version 1.0-4. Available at: http://CRAN.R-project.org/package=nortest.

Johnson, R.A. & Wichern, D.W. (2002) *Applied multivariate statistical analysis*. Upper Saddle River, NJ: Prentice Hall.

Lee, Y.S. (1971) Asymptotic formulae for the distribution of a multivariate test statistic: Power comparisons of certain multivariate tests. *Biometrika*, 58: 647-651.

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Pillai, K.C.S. & Jayachandran, K. (1967) Power comparisons of tests of two multivariate hypotheses based on four criteria. *Biometrika*, 54: 195-210.

Examples

```
## Not run:
data(ZXV3)

XVI1(data=ZXV3, var=c("Weaning.weight", "Birth.weight"), Factor=c("Father","Gender"))

## End(Not run)
```

XVI2 *Multivariate analysis of covariance (MANCOVA)*

Description

An analysis of covariance is applied considering several dependent quantitative variables.

Usage

```
XVI2(data, var, Factor, Covar, SS="III", trans=NULL, pthreshold=0.05, psig=FALSE,
ellipse=FALSE, convex=TRUE, PCH=rep(16,100), order=TRUE, cor=TRUE,
file1="Output.txt", file2="Multinormality.csv", file3="Post hoc pvalues.csv",
na="NA", dec=",", row.names=TRUE)
```

Arguments

`data`	Data file.
`var`	Dependent variables.
`Factor`	Variables with the factors.
`Covar`	Covariates.

`SS` When there are several factors, the decomposition of the sum of squares in parts attributed to each one of the factors is not unique. Four types of sums of squares are commonly used: type I for nested models, type II for balanced models, type III for unbalanced models (also balanced), and type IV for models with empty boxes. The type III is suitable for most applications, so it will be the one used by default. This description is very simplified; interested readers should consult supplementary bibliography for a correct use of the ANOVA models.

`trans` Type of transformation that is applied to the data:

1. NULL (untransformed)
2. 1/x2
3. 1/x
4. LN
5. LOG
6. SQR (square root)
7. x2
8. x3
9. EXP (exponential)
10. ASN (arcsine)

`pthreshold` Threshold probability that is used to group the categories of each factor, based on the probabilities obtained in the post hoc analysis.

`psig` If TRUE, in the post hoc inter-group comparison file, only those that are significant with a probability value equal to or less than the cut-off value specified in the argument *pthreshold* are shown.

`ellipse` If it is TRUE, the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the factors are depicted.

`convex` If it is TRUE, the convex hull is depicted for each category.

`PCH` Symbols of each category of the factors.

`order` If it is TRUE, the first variable of the polar coordinates is the one that best discriminates the categories. The objective is to show the greatest possible separation between the categories in the plots. If it is FALSE, the first variable is the one specified by the user in the argument *var*.

`cor` If TRUE, the variables are ordered according to the correlation between them when estimating the polar coordinates. Therefore, the next variable to another variable is the one with the highest positive correlation. The objective is to show the greatest possible separation between the categories in the plots. If it is FALSE and the argument *order=FALSE*, the variables are organized as they appear in the arguments *var* and *Covar*.

`file1` TXT FILE. Name of the output file with the results.

`file2` CSV FILE. Name of the file with the probabilities obtained from the multinormality analysis.

`file3` CSV FILE. Name of the file with the probabilities obtained from the post hoc comparison between groups for each factor.

`na` CSV FILE. Text that is used in the cells without data.

`dec` CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".

`row.names` CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

CONTRASTS OF HOMOGENEITY IN QUANTITATIVE VARIABLES

Multivariate analysis of covariance (MANCOVA)

The multivariate analysis of covariance (MANCOVA) is an extension of the univariate analysis of covariance (ANCOVA). Part of the variability observed in the dependent variables may be due to other variables. Therefore, if we really want to know if one or several factors explain the variability of dependent variables, it is necessary to eliminate any type of variation associated with other variables, which is done by introducing these variables as covariates. The reader is advised to read the *details* section of the IX4 function about ANCOVA, since, as mentioned above, the MANCOVA is an extension of the ANCOVA.

The assumptions of the MANCOVA are the same than as those of the MANOVA (see function XVI1 for more details), with the addition that the covariates must also be independent and there must be a linear relationship between the dependent variables and the covariates.

FUNCTIONS

The lillie.test function of the nortest package (Gross, 2015) is used to perform the Kolmogorov-Smirnov Normality test with the Lilliefors correction. The shapiro.test function is used to perform the Shapiro-Wilk test, the linear model is performed with the lm function, both of the base stats package. The MANCOVA type II-III is performed with the Manova function of the car package (Fox et al., 2018a). The MANCOVA type I is done with the function manova of the base stats package. To determine the homogeneity of variances between the categories of the factors, the Box's M test was applied, for which the function boxM of the heplots package (Fox et al., 2018b) was used. The function vif of the usdm package (Naimi, 2013; Naimi et al., 2014) was used for the VIF estimation.

EXAMPLE

Biometric data of two species of sharks (*Scyliorhinus canicula* and *Galeus melastomus*), which were taken by two researchers in four zones (two in the Mediterranean and two in the Atlantic).

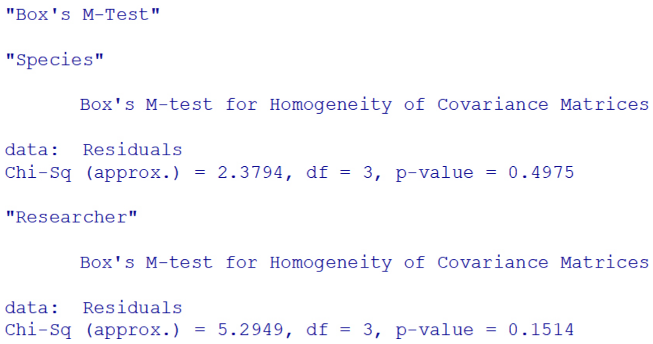
The aim is to determine if there are differences in the base of the anal fin and the height of the anal fin, between species and researchers, considering the effect of the total length of the shark.

In the script the argument *trans="SQR"* was included, that is, the dependent variables are transformed with the square root, in order to meet the requirements of multinormality and equality of the variances-covariances matrices in the residuals of the groups of factors.

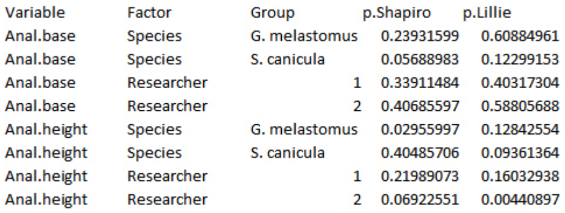
In the TXT file of the results, the first thing that is observed are the VIF values, which show that there is a bit of collinearity among the dependent variables. If the analysis is done without transforming the data, this collinearity is not observed, but the requirements of multinormality and homogeneity of variances are not met.

```
  Variables       VIF
  Anal.base   16.54237
Anal.height  16.54237
```

The assumption of equality of variances between groups of factors is also fulfilled in both factors: researcher (p = 0.151) and species (p = 0.497).



```
"Box's M-Test"

"Species"

       Box's M-test for Homogeneity of Covariance Matrices

data:  Residuals
Chi-Sq (approx.) = 2.3794, df = 3, p-value = 0.4975

"Researcher"

       Box's M-test for Homogeneity of Covariance Matrices

data:  Residuals
Chi-Sq (approx.) = 5.2949, df = 3, p-value = 0.1514
```



Variable	Factor	Group	p.Shapiro	p.Lillie
Anal.base	Species	G. melastomus	0.23931599	0.60884961
Anal.base	Species	S. canicula	0.05688983	0.12299153
Anal.base	Researcher	1	0.33911484	0.40317304
Anal.base	Researcher	2	0.40685597	0.58805688
Anal.height	Species	G. melastomus	0.02955997	0.12842554
Anal.height	Species	S. canicula	0.40485706	0.09361364
Anal.height	Researcher	1	0.21989073	0.16032938
Anal.height	Researcher	2	0.06922551	0.00440897

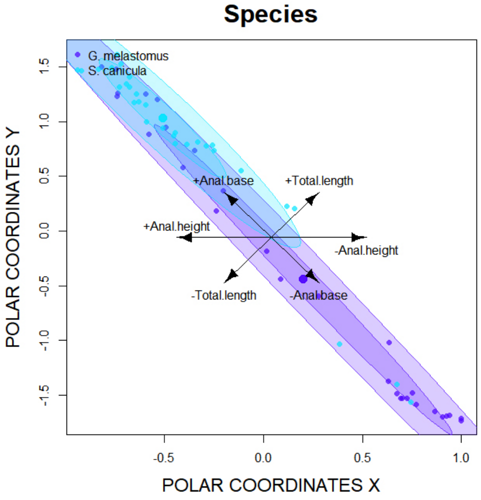
In the file that is called by default "Multinormality.csv", it is observed that the multinormality is also fulfilled (see table showed above). The samples are normal in both variables for all groups of the two factors, although in a couple of groups there are discrepancies between the two normality test.

The following results are the MANCOVA analysis. There are four tests that are used more frequently: Pillai, Wilks, Hotelling-Lawley and Roy. The following table shows that in all cases the test values coincide. The covariate total length is clearly related to the dependent variables (p <0.001). However, there are no significant differences between species (p = 0.207), nor among researchers (p = 0.534), nor are there any researcher-species interactions (p = 0.27). Non-interaction means that the two researchers measure the two species equally.

```
Multivariate Tests: Total.length
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            1   0.97544  1032.48      2     52 < 2.22e-16 ***
Wilks             1   0.02456  1032.48      2     52 < 2.22e-16 ***
Hotelling-Lawley  1  39.71078  1032.48      2     52 < 2.22e-16 ***
Roy               1  39.71078  1032.48      2     52 < 2.22e-16 ***
Multivariate Tests: Species
                 Df test stat approx F num Df den Df  Pr(>F)
Pillai            1 0.0587534 1.622942      2     52 0.20715
Wilks             1 0.9412466 1.622942      2     52 0.20715
Hotelling-Lawley  1 0.0624208 1.622942      2     52 0.20715
Roy               1 0.0624208 1.622942      2     52 0.20715
Multivariate Tests: Researcher
                 Df test stat  approx F num Df den Df  Pr(>F)
Pillai            1 0.0237820 0.6333962      2     52 0.53483
Wilks             1 0.9762180 0.6333962      2     52 0.53483
Hotelling-Lawley  1 0.0243614 0.6333962      2     52 0.53483
Roy               1 0.0243614 0.6333962      2     52 0.53483
Multivariate Tests: Species:Researcher
                 Df test stat approx F num Df den Df  Pr(>F)
Pillai            1 0.0490789 1.341911      2     52 0.27024
Wilks             1 0.9509211 1.341911      2     52 0.27024
Hotelling-Lawley  1 0.0516119 1.341911      2     52 0.27024
Roy               1 0.0516119 1.341911      2     52 0.27024
```

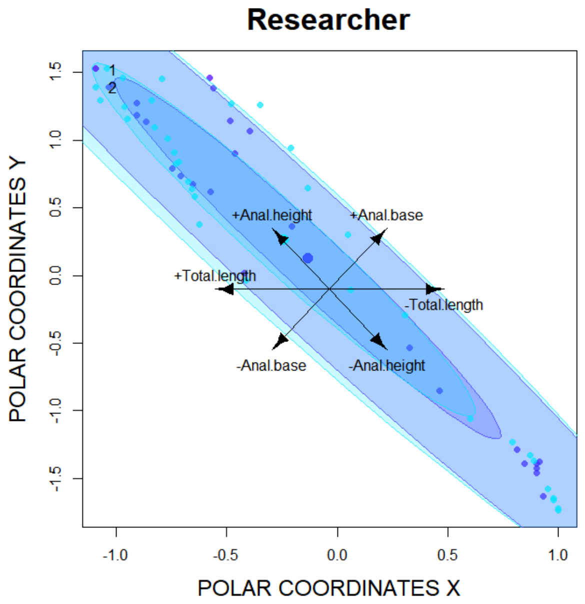
Figure XVI2.1 shows that there are no differences between species, which corroborates the MANCOVA results. It is important to mention that the plot includes the covariates, in order to better visualize the effect of them on the dependent variables.

Figure XVI2.1. Polar coordinates of the species as a function of the dependent variables and the covariate.



In the researcher (Figure XVI2.2), no differences are observed between them.

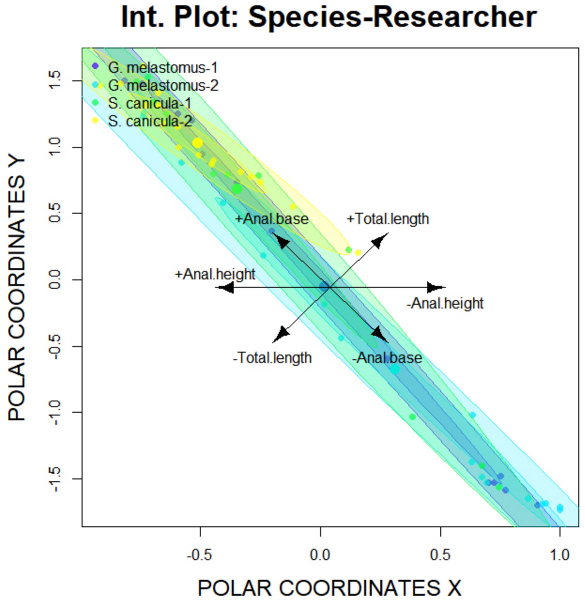
Figure XVI2.2. Polar coordinates of the species as a function of the dependent variables and the covariate.



There is no information about the post hoc tests nor the file with the comparison probabilities or the graphics, because in both factors there are only two groups.

The last plot is that of the interaction between both factors (Figure XVI2.3), where it is seen that both researchers measure both species equally and, therefore, there is no interaction.

Figure XVI2.3. Polar coordinates of the interaction between both factors



All these polar coordinate graphs can be modified to suit the user using the original data and the F46 function to perform polar coordinates available in the PlotsR package.

Value

It is exported:

1. A TXT file with the results of the VIF, the Box' M test and the results of the MANCOVA.

2. A CSV file with the probabilities obtained from the multinormality analysis.

3. A CSV file with the probabilities obtained from the post hoc comparison between the groups of each factor.

4. Polar coordinate plots showing the separation between categories for each factor, the groups resulting from the post hoc analysis and the interactions between factors.

References

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2018a) Companion to Applied Regression. R package version 3.0-0. Available at: http://CRAN.R-project.org/package=car.

Fox, J, Friendly, M. & Monette, G. (2108b) Visualizing Hypothesis Tests in Multivariate Linear Models. R package version 1.3.5. Available at: http://CRAN.R-project.org/package=heplots.

Gross, J. (2015) Tests for Normality. R package version 1.0-4. Available at: http://CRAN.R-project.org/package=nortest.

Naimi, B. (2013) Uncertainty analysis for species distribution models. R package version 3.5-0. Available at: http://CRAN.R-project.org/package=usdm.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37: 191-203.

Examples

```
## Not run:
data(ZIX3)

XVI2(data=ZIX3, var=c("Anal.base", "Anal.height"), Factor=c("Species",
"Researcher"), Covar="Total.length", trans="SQR", ellipse=TRUE, convex=FALSE)

## End(Not run)
```

ZII1 *FISH ABUNDANCE*

Description

Abundance of various fish species in two areas of sampling (S1 and S2), and temperature and salinity in each of these areas.

Usage

```
data(ZII1)
```

Format

A data frame with 5 columns including the taxonomy of several species of fish, 2 columns of abundance data in two zones, and temperature and salinity in both zones.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZIV1 *HEIGHT AND WEIGHT*

Description

Height and weight, in different countries and cities, of men and women.

Usage

```
data(ZIV1)
```

Format

A data frame with 5 columns: country, city, sex, height (in cm) and weight (in kg).

ZIV3 *DURATION OF IMMUNITY*

Description

The length of time (in months) that a group of women and men remain immune after they are given a vaccine.

Usage

```
data(ZIV3)
```

Format

A data frame with 2 columns: duration of immunity and sex.

ZIX1 *HEIGHT AND WEIGHT*

Description

Height and weight data for children aged 2-4 years in Italy and Spain.

Usage

```
data(ZIX1)
```

Format

A data frame with 5 columns: country, gender, age,weight and height.

ZIX10 *ERYTHROCYTES CONTAINED IN BLOOD IN DIFFERENT INDIVIDUALS*

Description

Data on the concentration of erythrocytes (in millions per cubic millimeter) of several men and women who underwent treatment for increasing the concentration of red blood cells. The difference here is that the file ZIX9 identifies each person with a tag.

Usage

```
data(ZIX10)
```

Format

A data frame with 5 columns: individual and erythrocyte concentrations at the beginning and for the three consecutive months (months 1, 2 and 3) after starting treatment.

ZIX11 *CLIMATE DATA*

Description

Daily climate data for 1990-2000 in three cities in Spain: Huelva, Palma de Mallorca and Vigo.

Usage

```
data(ZIX11)
```

Format

A data frame with 14 columns: city, altitude, year, month, day, high temperature, low temperature, average temperature, wind direction, wind speed, precipitation, sun, high pressure and low pressure.

Source

http://www.aemet.es/es/portada

ZIX3 *SHARK MORPHOMETRY*

Description

Biometric data of two species of sharks(*Scyliorhinus canicula* and *Galeus melastomus*), which were taken by two researchers in four areas (two in the Mediterranean and two in the Atlantic).

Usage

```
data(ZIX3)
```

Format

A data frame with 6 columns: species, researcher, total length, length of the base of the anal fin and the anal-fin height.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZIX5 *ABILITY TO REMEMBER*

Description

In an experiment conducted with expert tasters and people who had no experience tasting, they were taught to identify 15 types of wines from different regions. Variations in ability to ascertain the wine provenance over time (after one hour, one day, one week and one month) was measured between experts and non-experts and, if the people's age affected the ability to succeed. For every time, each person assessed a large number of samples and the degree of success was recorded on a scale of 0 to 12.

Usage

```
data(ZIX5)
```

Format

A data frame with 7 columns: taster, age, if he/she has or does not have any experience (YES / NO), and the measurement time (Hour, Day, Week and Month), with a degree of success on a scale of 0 to 12.

ZIX6 *LICHENS AS INDICATORS OF POLLUTION*

Description

Average data of lichen species found in two species of tree, *Populus alba* (White Poplar) and *Aesculus hippocastanum* (Indian horse chestnut), in three cities in (Madrid, Barcelona and Seville), in three types of environments (low, medium and high pollution) and measured by two researchers.

Usage

```
data(ZIX6)
```

Format

A data frame with 5 columns: city, tree species, degree of contamination, researchers and richness of lichen species.

ZIX7 *NUTRIENTS IN AMAZON LAKES*

Description

Data of the nitrate concentration, nitrate and ammonium (microM/L) in lakes of the Colombian Amazon in two different months.

Usage

```
data(ZIX7)
```

Format

A data frame with 5 columns: month, lake and concentrations of nitrite, nitrate and ammonium.

ZIX9 *ERYTHROCYTES CONTAINED IN BLOOD ACCORDING TO SEX*

Description

Data on the concentration of erythrocytes (in millions per cubic millimeter) of several men and women who underwent treatment for increasing the concentration of red blood cells.

Usage

```
data(ZIX9)
```

Format

A data frame with 5 columns: sex and erythrocyte concentrations at baseline and in the months 1, 2 and 3 after starting treatment.

ZVI1 *ABUNDANCE OF SEA SNAILS*

Description

Abundance of two species of herbivorous sea snails(*Littorina littorea* and *Littorina saxatilis*) and a carnivorous species (*Nucella lapillus*) in two sampling areas.

Usage

```
data(ZVI1)
```

Format

A data frame with 4 columns: genus, species and abundance of species in two sampling areas.

ZVII1 *STUDY ON SMOKING*

Description

Range data in men and women who smoke in different work centres. The categories used were: 1 (Non-smoker), 2 (between 1 and 10 cigarettes a day), 3 (between 11 and 20 cigarettes a day), 4 (from 1 to 2 packs per day) and 5 (more than 2 packs a day). There is also information if any parents of these workers are smokers and their categories are: workers in which one parent is a smoker (category value = 1) and the other group for those in which none of his/her parents is a smoker (category value = 0).

Usage

```
data(ZVII1)
```

Format

A data frame with 4 columns: gender, workplace, if either parent smokes and degree of smoking.

ZVII2 *THE EFFECTS OF HERPES SIMPLEX LABIALIS VIRUS*

Description

Data on the presence of herpes simplex labialis virus (0 is a non-viral carrier and 1 is a viral carrier) in men and women (1 is man and 2 is woman).

Usage

```
data(ZVII2)
```

Format

A data frame with 2 columns: sex and whether the person is or is not a carrier of the virus.

ZVII3 *WINE TASTING*

Description

Data on how 15 Enologysts test 4 different types of wine. Category 1 is very bad, Category 2 is bad, Category 3 is good and Category 4 is very good.

Usage

```
data(ZVII3)
```

Format

A data frame with 3 columns: enologyst, wine type and quality.

ZVII4 *SKIN ALLERGIC REACTION TO ONE ANTIBIOTIC IN TWO CONTROL SESSIONS*

Description

An antibiotic was given to 50 people and 50 others were given a placebo. The results compare the positive response (allergy) or negative (no allergy) in 100 people between days 5 and 10 after treatment.

Usage

```
data(ZVII4)
```

Format

A data frame with 3 columns: people who were provided and were not provided the antibiotic, response to the 5th and 10th day.

ZVII5 *ALLERGIC SKIN REACTION TO AN ANTIBIOTIC APPLIED IN DIFFERENT CONTROL SESSIONS*

Description

An antibiotic was given to 50 people and 50 others were given a placebo. The results compare the positive response (allergy) or negative (no allergy) in 100 people between 5, 10, 15 and 20 days after treatment.

Usage

```
data(ZVII5)
```

Format

A data frame with 5 columns: people who were given and were not given an antibiotic, and response to the 5th, 10th, 15th and 20th day.

ZVIII1 *ALLERGIC SKIN REACTION TO AN ANTIBIOTIC TESTED OVER TIME*

Description

An antibiotic was given to 50 people and 50 others were given a placebo. The results compare the positive response (allergy) or negative (no allergy) in 100 people between 5, 10, 15 and 20 days after treatment.

Usage

```
data(ZVIII1)
```

Format

A data frame with 3 columns: people who were given and were not given an antibiotic, and the type of response to the 5th, 10th, 15th and 20th days.

ZVIII2 *INFANT MORTALITY*

Description

Data from a sample of infants with birth weight: I (less than 1500 g), II (from 1500 to 2499 g), III (from 2500 to 4199 g) and IV (more than 4200 g) in two hospitals and whether the baby died before his/her first year of life.

Usage

```
data(ZVIII2)
```

Format

A data frame with 3 columns: hospital, group in which the baby was according to his/her birth weight and whether the baby lived or died before the first year of life.

ZX1 *QUAIL BREEDING*

Description

Data collection and sampling made in different provinces and towns in Spain, of clutches from a small bird species, common quail.

Usage

```
data(ZX1)
```

Format

A data frame with 6 columns: province, city, number of eggs per female, female age, average temperature in the area during the breeding season, and food availability for females (whose values are percentages of the maximum measured).

ZX3 *MORPHOMETRIC VARIABLES IN CHARACIFORMS*

Description

Morphometric data of several species of Characiforms, as the length of the dorsal fin base (M12), body height (M11), etc. For details see Guisande et al. (2010).

Usage

```
data(ZX3)
```

Format

A data frame with 31 columns: taxonomic data (order, family, genus and species) and 26 morphometric variables.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZXI1 *MORPHOMETRY OF SHELLFISH*

Description

Morphometric variables of the shell of several species of molluscs.

Usage

```
data(ZXI1)
```

Format

A data frame with 5 columns: species, weight (in grams) and length, height and width of the shell (in cm).

ZXI10 *FISHES IDENTIFICATION BASED ON THEIR MORPHOMETRY-2*

Description

Data of body measurements of freshwater fish. These data are used to determine the predictability, the multinomial logit model, when identifying species, since they were not used in the example of the function XI8.

Usage

```
data(ZXI10)
```

Format

A data frame with 15 columns, where the first three are the taxonomy of several species of freshwater fish and the next 12 columns are body measurements.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZXI11 *EXPOSURE TO ARSENIC AND MORTALITY FROM RESPIRATORY DISEASES-1*

Description

Dataset from a cohort study on arsenic exposure in industry and deaths from respiratory diseases, available in the database Montana of old package *epicalc* (Chongsuvivatwong, 2012).

Usage

```
data(ZXI11)
```

Format

A data frame with 7 columns: number of deaths in each group (*respdeath*), number of people per year of exposure (*personyrs*), dividing the number of deaths and the number of people by years of exposure * 10,000 (*respdeath.personyrs*), age groups (*agegr*), period of employment (*period*), commencement of employment (*start*) and time of exposure to arsenic in years (*arsenic*).

References

Chongsuvivatwong, V. (2012) Epidemiological calculator. R package version 2.15.1.0.

ZXI12 *EXPOSURE TO ARSENIC AND MORTALITY FROM RESPIRATORY DISEASES-2*

Description

Dataset from a cohort study on arsenic exposure in industry and deaths from respiratory diseases, available in the database Montana of the old package *epicalc* (Chongsuvivatwong, 2012). These data are used to determine the predictive power of Poisson regression estimated in the example of the function XI10, because they were not used for the model estimation.

Usage

```
data(ZXI12)
```

Format

A data frame with 7 columns: number of deaths in each group (*respdeath*), number of people per year of exposure (*personyrs*), dividing the number of deaths and the number of people by years of exposure * 10,000 (*respdeath.personyrs*), age groups (*agegr*), period of employment (*period*), commencement of employment (*start*) and time of exposure to arsenic in years (*arsenic*).

References

Chongsuvivatwong, V. (2012) Epidemiological calculator. R package version 2.15.1.0.

ZXI2 *WEALTH ACCUMULATED OF SPECIES*

Description

This data comes from a study in which different sampling sites and the total cumulative number of species are recorded the sampled sites.

Usage

```
data(ZXI2)
```

Format

A data frame with 2 columns with the sampling site and the accumulated wealth including all the previously sampled sites.

`ZXI3` *TOXIC PHYTOPLANKTON GROWTH*

Description

Experiment in which the growth rate of a species of toxic phytoplankton genus *Alexandrium* based on the phosphate concentration of the culture was measured.

Usage

```
data(ZXI3)
```

Format

A data frame with 2 columns: growth rate d^{-1} and phosphate concentration μM.

References

Frangópulos, M., Guisande, C., deBlas, E. & Maneiro, I. (2004) Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae* 3: 131-139.

`ZXI4` *ANCHOVY GROWTH*

Description

Relationship between length and age of the individual, quantified by counting otolith rings in anchovy (*Engraulis encrasicolus*), in the Alboran Sea and the Strait of Sicily.

Usage

```
data(ZXI4)
```

Format

A data frame with 3 columns. Age (in days), length (in cm) and area.

References

Basilone, G., Guisande, C., Patti, B., Mazzola, S., Cuttitta, A., Bonanno, A. & Kallianiotis, A. (2004). Linking habitat conditions and growth in the European anchovy (*Engraulis encrasicolus*). *Fisheries Research*, 68: 9-19.

ZXI5 *PHYTOPLANKTON ABUNDANCE*

Description

Evolution of the abundance of phytoplankton species in culture conditions over time.

Usage

```
data(ZXI5)
```

Format

A matrix with 2 columns: days of culture and abundance in cells per milliliter.

References

Frangópulos, M., Guisande, C., deBlas, E. & Maneiro, I. (2004) Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae* 3: 131-139.

ZXI6 *PRIMARY PRODUCTION IN UPWELLING AREAS*

Description

Chlorophyll data, upwelling, stability, platform size, temperature and nutrients from different coastal areas of the world.

Usage

```
data(ZXI6)
```

Format

A data frame with 12 columns: upwelling area, latitude, longitude, chlorophyll concentration (mg^{-3}), percentage of continental shelf in the cell, Ekman transport $(m^3km^{-1}s^{-1})$ which is a measure of upwelling, temperature, turbulence (m^3s^{-1}), stability of the water column$(cycless^{-1})$ and concentrations of phosphates, nitrates and silicates in μMl^{-1}.

References

Patti, B., Guisande, C., Vergara, A.R., Riveiro, I., Maneiro, I., Barreiro, A., Bonanno, A., Buscaino, G., Cuttitta, A., Basilone, G. & Mazzola, S. (2008) Factors responsible for the differences in satellite-based chlorophyll a concentration between the major global upwelling areas. *Estuarine, Coastal and Shelf Science*, 76, 775-786.

ZXI7 *CANCER PATIENTS-1*

Description

Data from patients who have been diagnosed with a type of cancer as well as people who have not been diagnosed.

Usage

```
data(ZXI7)
```

Format

A data frame with 5 columns: if the person is or is not suffering from cancer, age, sex, marital status, and whether blood in urine has been detected.

ZXI8 *CANCER PATIENTS-2*

Description

Data from patients who have been diagnosed with a type of cancer, as well as people who do not. This data is used to determine the predictive power of the binomial logistic model estimated in the example of the function XI6, because they were not used to estimate the model.

Usage

```
data(ZXI8)
```

Format

A data frame with 5 columns: if the person is or is not suffering from cancer, age, sex, marital status, and whether blood in urine has been detected.

ZXI9 *FISHES IDENTIFICATION BASED ON THEIR MORPHOMETRY-1*

Description

Data of body measurement of freshwater fish, which are used to identify the species using logistic regression as a statistical technique.

Usage

```
data(ZXI9)
```

Format

A data frame with 15 columns, where the first three are the taxonomy of several species of freshwater fish and the next 12 columns are body measurements.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZXII1 *INFLUENCE OF MEDICAL ADVICE TO QUIT SMOKING*

Description

Data refer to a study by Silagy & Ketteridge (1997) on the effect of medical advice when quitting smoking. In the data file, the variable *event.e* refers to the number of people who quit smoking by doctor's recommendation, *n.e* the total number of people in the group to whom the doctor recommended to stop smoking and, *event.c* and *n.c* are the people who quit smoking and the number of people who were in the control group, to whom the doctor advised not to quit smoking, respectively.

Usage

```
data(ZXII1)
```

Format

A data frame with 5 columns: study, people who quit smoking by doctor's recommendation (*event.e*), total number of people to whom the doctor recommended stop smoking(*n.e*), total number of people who quit smoking in the control group(*event.c*) and total number of people in the control group (*n.c*). In the control group the doctor did not recommend anyone to stop smoking.

References

Silagy, C. & Ketteridge, S. (1977) Physician advice for smoking cessation. In: Lancaster T, Silagy C, Fullerton D. Editores. Tobacco Addiction Module of the Cochrane Database of Systematic Reviews 4.

`ZXIII1` *ASSESSMENT CRITERIA FOR JOB CANDIDATES*

Description

Assessment criteria, on a scale from 0 to 10, of 48 candidates vying for a job.

Usage

```
data(ZXIII1)
```

Format

A data frame of 15 columns with a candidate code in the first column and the assessment criteria which are: letter of motivation, level of studies, sympathy, self-confidence, lucidity, honesty, business sense, experience, charisma, ambition, understanding, potential, motivation for the post and adaptation.

`ZXIII2` *DEMOGRAPHIC PARAMETERS FOR COUNTRIES*

Description

Demographic parameters from 57 countries in Europe, Africa and America.

Usage

```
data(ZXIII2)
```

Format

A data frame with 11 columns: continent, country, male and female expectancy at birth (years of life), death rates, infant mortality, birth and fertility, the gross domestic product per capita (thousands of dollars each year) and the rate of literacy among men and women (in percentage) in the year 2000. The data were obtained from The World Bank (`http://www.worldbank.org/`).

ZXIII3 *MORPHOMETRY OF SHARK FAMILIES*

Description

Biometric data of several species belonging to four families of sharks.

Usage

```
data(ZXIII3)
```

Format

A data frame with 24 columns: the family and the rest of the columns are morphometric measurements.

Source

http://www.ipez.es/ipez/index_country/index.html

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

ZXIII4 *SOIL QUALITY AND REPRODUCTIVE GROWTH PARAMETERS IN PLANTS*

Description

Data on soil quality and parameters that measure growth and reproduction in plants.

Usage

```
data(ZXIII4)
```

Format

A data frame with 9 columns: the sampling site, a group of variables that gives information on soil quality: water content (percentage of soil weight), nitrogen, phosphate, and a number of trace minerals (in micromoles per gram of soil), and another set of variables which provides an idea on condition of plants, and those are a series of independent parameters related to growth and reproduction: foliar coverage (m^2), the average internode length (in cm) and the percent of the total viable seed production.

ZXIV1 *DEMOGRAPHIC PARAMETERS OF DIFFERENT REGIONS OF SPAIN*

Description

Demographic data presented for 2010 of the 19 regions or Autonomous Communities of Spain published by the National Statistics Institute (http://www.ine.es).

Usage

```
data(ZXIV1)
```

Format

A data frame with 10 columns: region, state, number of children per mother, average age when woman has her first child, average percentage of unmarried women with children, life expectancy for men and women, and the number of births , population mortality and child mortality (under 1 year), the last three demographic parameters per 1000 inhabitants.

Source

http://www.ine.es

ZXIV2 *NUTRIENTS AND ABUNDANCE OF PHYTOPLANKTON IN LAKES OF COLOMBIA*

Description

Relationship between nutrient concentration and phytoplankton abundance, in several sampling stations in different lakes in Colombia.

Usage

```
data(ZXIV2)
```

Format

A data frame of 15 columns: region of Colombia, lake, nutrient concentrations and the abundance of various groups of phytoplankton.

ZXIV3 *FLOW CYTOMETRIC DATA*

Description

Morphological characteristics of phytoplankton species obtained from a flow cytometer.

Usage

```
data(ZXIV3)
```

Format

A data frame with 8 columns: species, roughness, size and emission at different wavelengths.

ZXIV4 *ALGAE PHYTOPLANKTON PIGMENTS*

Description

Relative concentrations of 19 pigments regarding the concentration of chlorophyll of phytoplankton species belonging to different kinds of algae.

Usage

```
data(ZXIV4)
```

Format

A data frame with 21 columns: class, species and 19 pigments.

ZXV1 *ANCHOVY CONDITION FACTOR*

Description

Monthly average values of the anchovy condition factor (*Engraulis encrasicolus*), temperature (in degree C), chlorophyll (in mgm^{-3}) and stability of the water column (in m^3s^{-3}) over several years in the Strait of Sicily (Basilone et al., 2006).

Usage

```
data(ZXV1)
```

Format

A data frame with 6 columns: year, month, condition factor, chlorophyll, temperature and stability of the water column.

References

Basilone, G., Guisande, C., Patti, B., Mazzola, S., Cuttitta, A., Bonanno, A., Vergara, A.R. & Maneiro, I. (2006) Effect of habitat conditions on reproduction of the European anchovy (*Engraulis encrasicolus*) in the Strait of Sicily. *Fisheries Oceanography*, 15: 271-280.

`ZXV2` *TIDAL RANGE AND ANGLE OF THE SUN AND THE MOON*

Description

Angles between the Sun and the Moon and difference in meters between high tide and low tide to 42º 8´ North and 14º 13´ West from 1/1/2004 to 14/3/2004.

Usage

```
data(ZXV2)
```

Format

A data frame with 3 columns: date, angle between the Sun and the Moon, and difference in meters between high tide and low tide.

`ZXV3` *WEIGHT OF CALVES*

Description

Weight of calves at birth and at weaning from different fathers and mothers.

Usage

```
data(ZXV3)
```

Index